Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities

Hima Lakkaraju

HARVARD

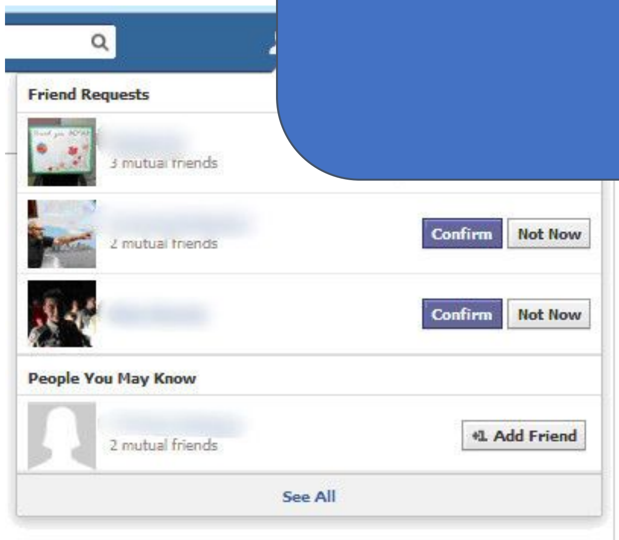
Schedule for Today

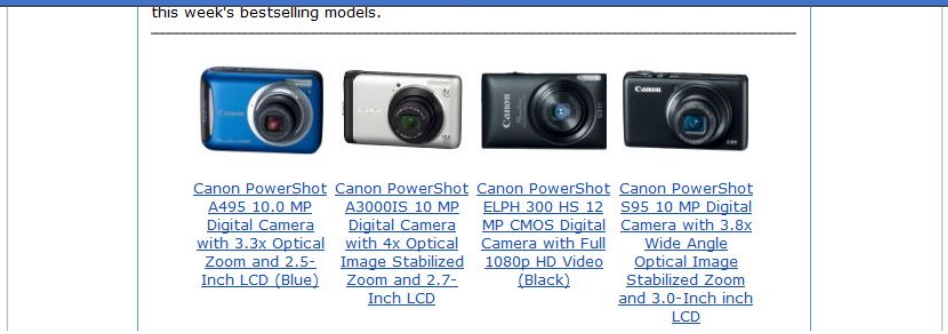
- **9am to 1020am**: Introduction; Overview of Inherently Interpretable Models
- **1020am to 1040am**: Break
- **1040am to 12pm**: Overview of Post hoc Explanation Methods
- **12pm to 1pm**: Lunch
- **105pm to 125pm**: Breakout Groups
- **125pm to 245pm**: Evaluating and Analyzing Model Interpretations and Explanations
- **245pm to 3pm**: Break
- **3pm to 4pm:** Analyzing Model Interpretations and Explanations, and Future Research Directions

[Weller 2017]

Motivation

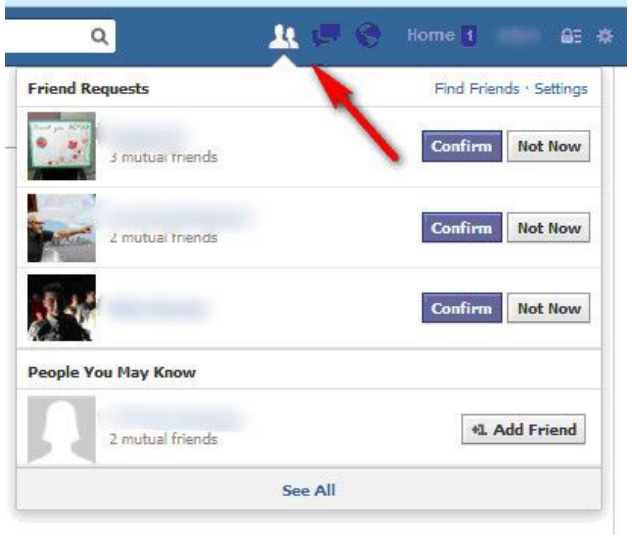
Machine Learning is EVERYWHERE!!

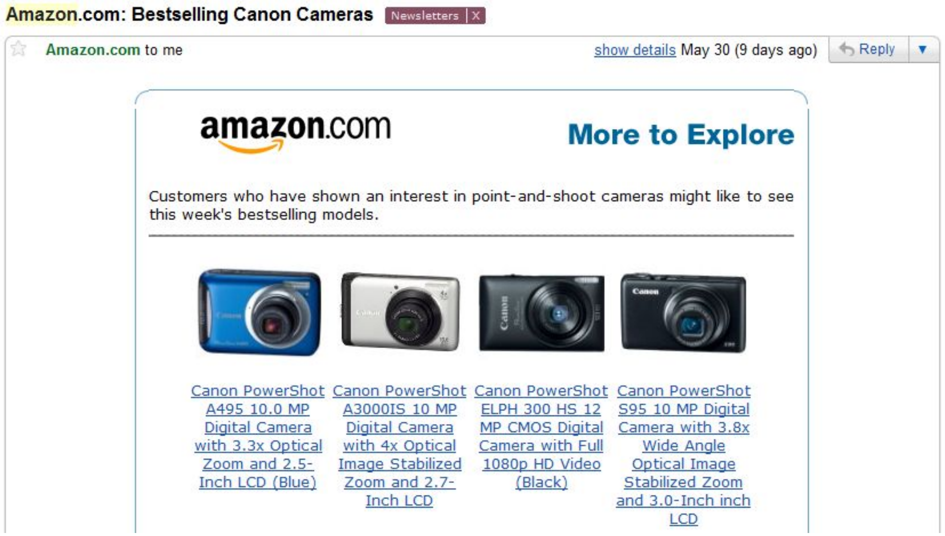


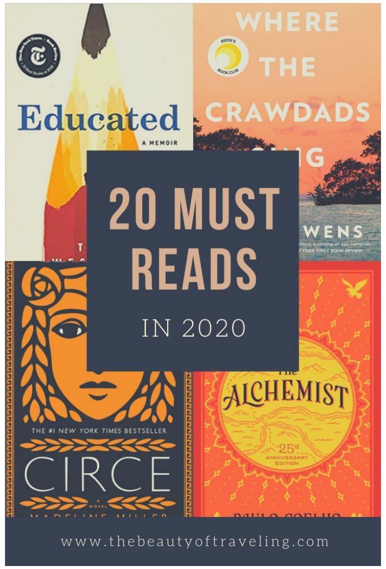


[Weller 2017]

Is Model Understanding Needed Everywhere?







[Weller 2017, Lipton 2017, Doshi-Velez and Kim 2016]

When and Why Model Understanding?

- Not all applications require model understanding
  - E.g., ad/product/friend recommendations
  - No human intervention

- Model understanding not needed because:
  - Little to no consequences for incorrect predictions
  - Problem is well studied and models are extensively validated in real-world applications trust model predictions

When and Why Model Understanding?

ML is increasingly being employed in complex high-stakes settings

When and Why Model Understanding?

- High-stakes decision-making settings
  - Impact on human lives/health/finances
  - Settings relatively less well studied, models not extensively validated
- Accuracy alone is no longer enough
  - Train/test data may not be representative of data encountered in practice
- Auxiliary criteria are also critical:
  - Nondiscrimination
  - Right to explanation
  - Safety

When and Why Model Understanding?

- Auxiliary criteria are often hard to quantify (completely)
  - E.g.: Impossible to predict/enumerate all scenarios violating safety of an autonomous car

- *Incompleteness* in problem formalization
  - Hinders optimization and evaluation
  - Incompleteness $\neq$ Uncertainty; Uncertainty can be quantified

When and Why Model Understanding?

Model understanding becomes critical when:

(i) models not extensively validated in applications; train/test data not representative of real time data

(ii) key criteria are hard to quantify, and we need to rely on a "you will know it when you see it" approach

Example: Why Model Understanding?

Input

Predictive Model

Prediction = Siberian Husky

This model is
incorrect
make
on!! Let
model

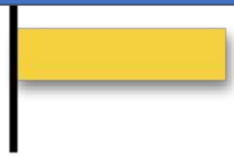
Model understanding facilitates debugging.

Example: Why Model Understanding?

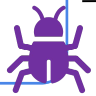
This prediction is

Defendant Detai

Model understanding facilitates bias detection.

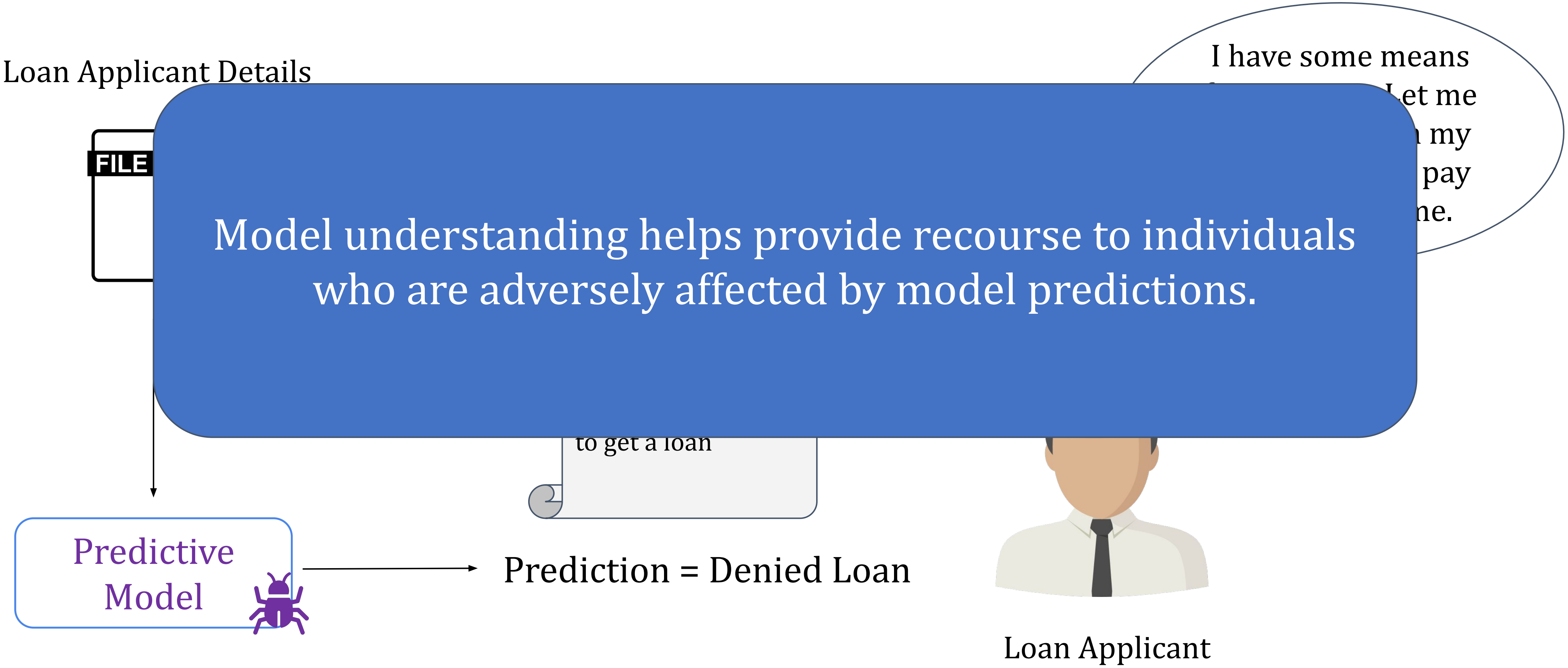


Predictive Model



Prediction = Risky to Release

Example: Why Model Understanding?



Example: Why Model Understanding?

Patient Data

25, Femal
32, Male,
31, Male,
.
.
.
.

Model Understanding

This model is using
irrelevant features when
on female
. I should
edictions
oup.

Model understanding helps assess if and when to trust model predictions when making decisions.

Predictive Model

Sick
Sick
.
.
Healthy
Healthy
Sick

Example: Why Model Understanding?

Model understanding allows us to vet models to determine if they are suitable for deployment in real world.

Summary: Why Model Understanding?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

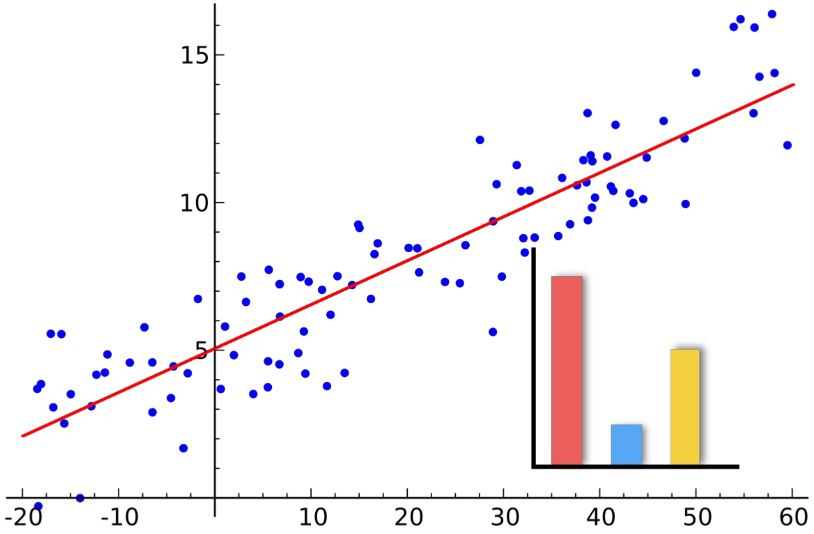
Decision makers (e.g., doctors, judges)

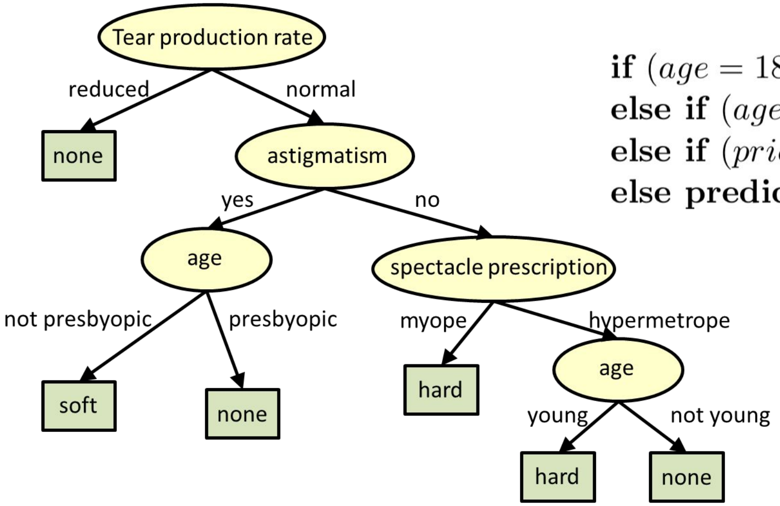
Regulatory agencies (e.g., FDA, European commission)

Researchers and engineers

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models

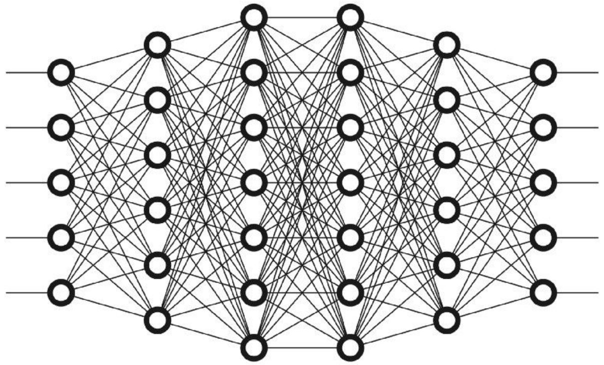




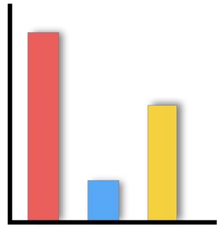
if $(age = 18-20)$ **and** $(sex = male)$ **then predict** *yes*
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** *yes*
else if $(priors > 3)$ **then predict** *yes*
else predict *no*

Achieving Model Understanding

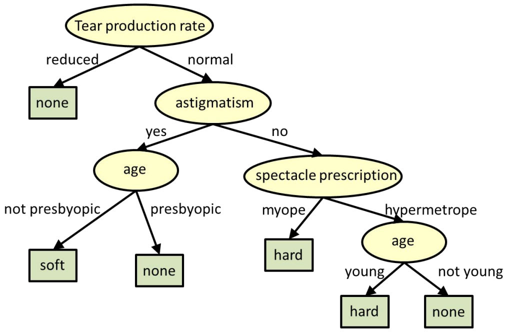
Take 2: *Explain* pre-built models *in a post-hoc manner*



Explainer

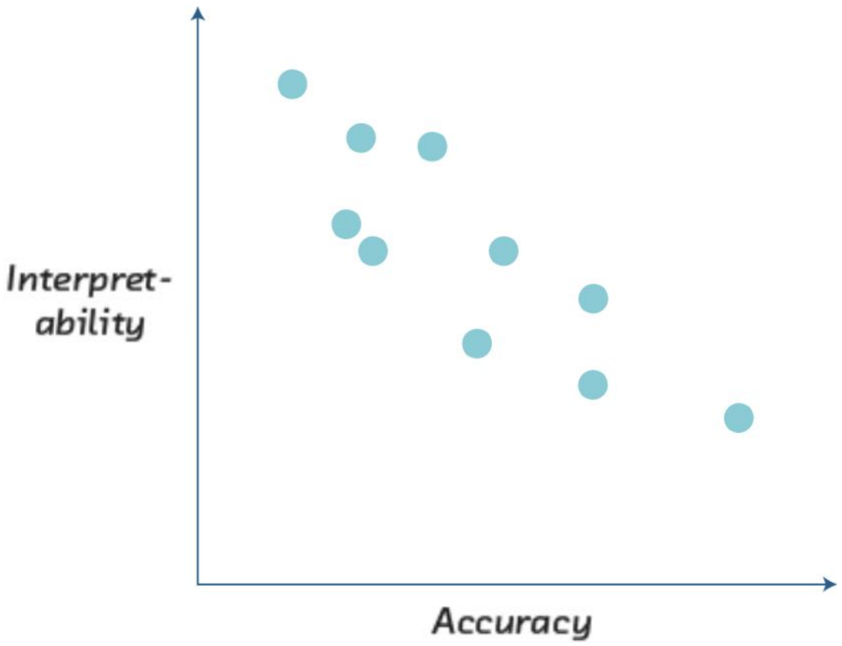


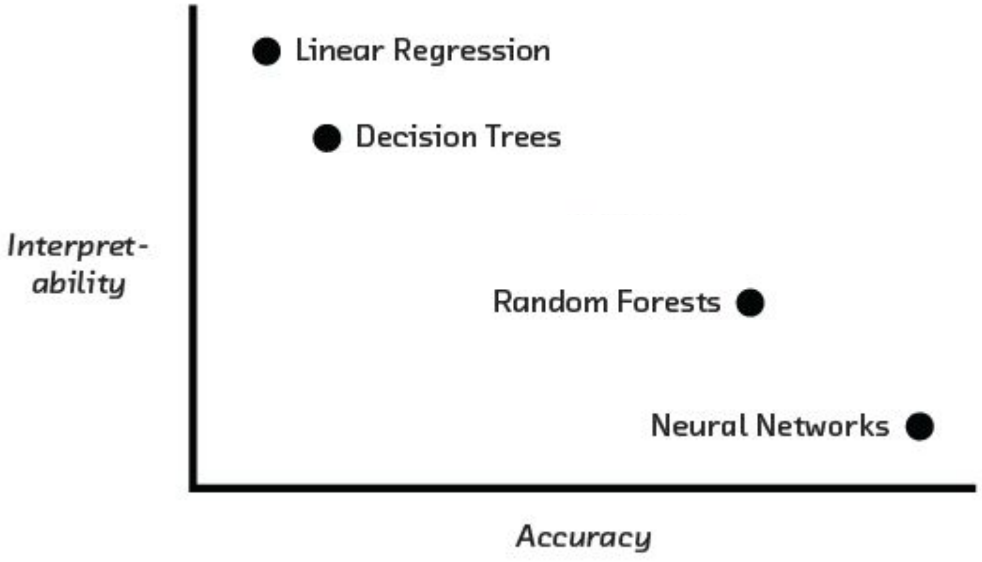
if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no



Inherently Interpretable Models vs. Post hoc Explanations

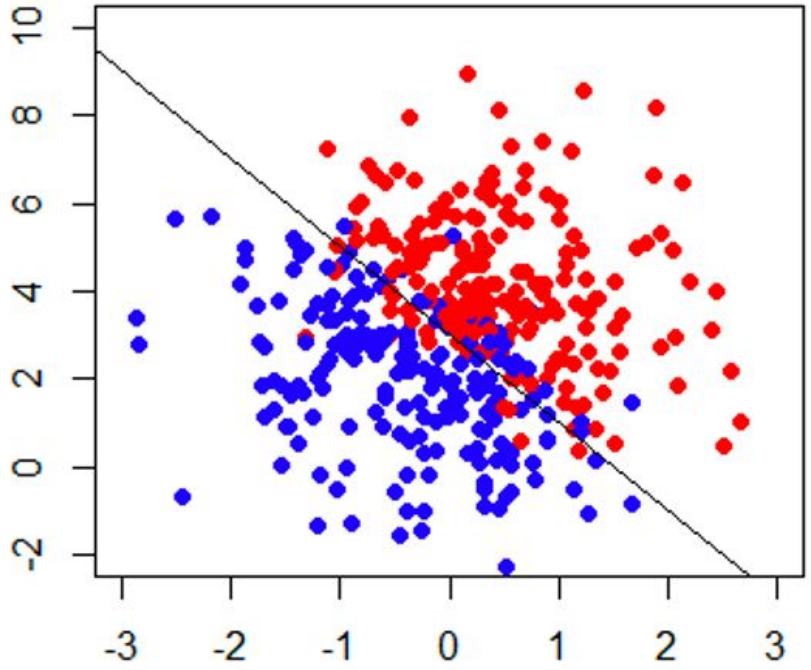
Example



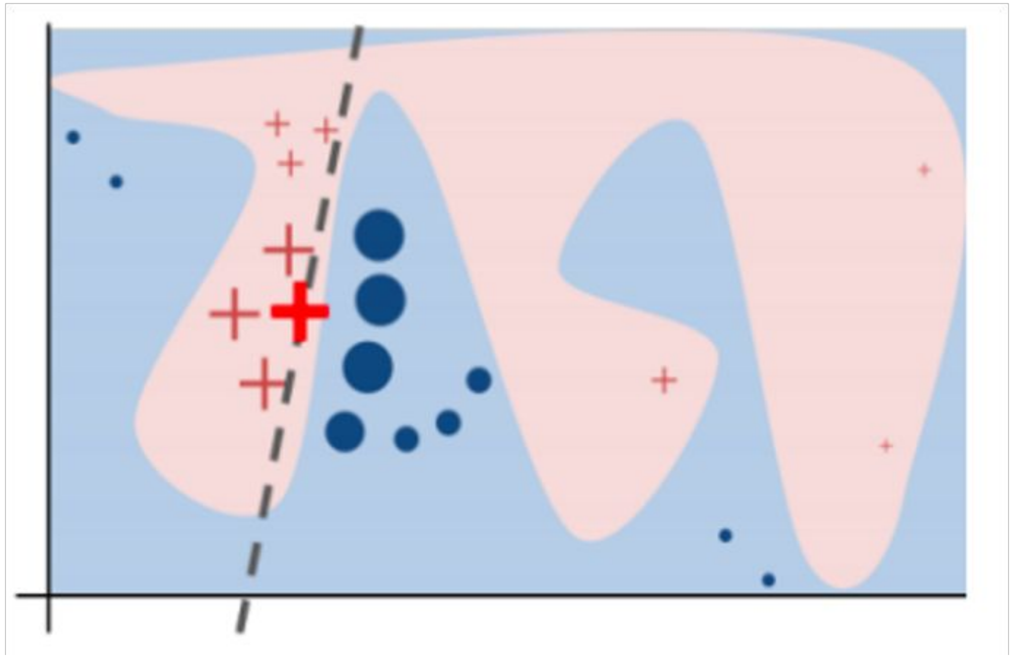


In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

Inherently Interpretable Models vs. Post hoc Explanations



can build interpretable + accurate models



complex models might achieve higher accuracy

[Ribeiro et. al. 2016]

Inherently Interpretable Models vs. Post hoc Explanations

Sometimes, you don't have enough data to build your model from scratch.

And, all you have is a (proprietary) black box!

Inherently Interpretable Models vs. Post hoc Explanations

If you *can build* an ***interpretable model*** which is also adequately accurate for your setting, DO IT!

Otherwise, ***post hoc explanations*** come to the rescue!

Agenda

- ❑ Inherently Interpretable Models
- ❑ Post hoc Explanation Methods
- ❑ Evaluating Model Interpretations/Explanations
- ❑ Empirically & Theoretically Analyzing Interpretations/Explanations
- ❑ Future of Model Understanding

Agenda

- ❑ Inherently Interpretable Models
- ❑ Post hoc Explanation Methods
- ❑ Evaluating Model Interpretations/Explanations
- ❑ Empirically & Theoretically Analyzing Interpretations/Explanations
- ❑ Future of Model Understanding

Inherently Interpretable Models

- Rule Based Models
- Risk Scores
- Generalized Additive Models
- Prototype Based Models
- Attention Based Models

Inherently Interpretable Models

- Rule Based Models
- Risk Scores
- Generalized Additive Models
- Prototype Based Models
- Attention Based Models

[Letham et. al. 2016]

Bayesian Rule Lists

- A rule list classifier for stroke prediction

if hemiplegia **and** age > 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age > 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age ≤ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

Bayesian Rule Lists

- A generative model designed to produce rule lists (if/else-if) that strike a balance between accuracy, interpretability, and computation

- What about using other similar models?
  - Decision trees (CART, C5.0 etc.)
  - They employ greedy construction methods
  - Not computationally demanding but affects quality of solution – both accuracy and interpretability

Bayesian Rule Lists: Generative Model

– Sample a decision list length $m \sim p(m|\lambda)$.
– Sample the default rule parameter $\theta_0 \sim \text{Dirichlet}(\alpha)$.
– For decision list rule $j = 1, \ldots, m$:
  Sample the cardinality of antecedent a_j in d as $c_j \sim p(c_j|c_{<j}, \mathcal{A}, \eta)$.
  Sample a_j of cardinality c_j from $p(a_j|a_{<j}, c_j, \mathcal{A})$.
  Sample rule consequent parameter $\theta_j \sim \text{Dirichlet}(\alpha)$.
– For observation $i = 1, \ldots, n$:
  Find the antecedent a_j in d that is the first that applies to x_i.
  If no antecedents in d apply, set $j = 0$.
  Sample $y_i \sim \text{Multinomial}(\theta_j)$.

$\mathcal{A}$ is a set of pre-mined antecedents

Model parameters are inferred using the Metropolis-Hastings algorithm which is a Markov Chain Monte Carlo (MCMC) Sampling method

Pre-mined Antecedents

- A major source of practical feasibility: pre-mined antecedents
  - Reduces model space
  - Complexity of problem depends on number of pre-mined antecedents

- As long as pre-mined set is expressive, accurate decision list can be found + smaller model space means better generalization (Vapnik, 1995)

[Lakkaraju et. al. 2016]

Interpretable Decision Sets

- A decision set classifier for disease diagnosis

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure$\geq$ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI$\geq$ 0.3 **and** Insurance=None **and** Blood-Pressure$\geq$ 0.2 **then** Depression

If Smoker=Yes **and** BMI$\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

If Risk-Diabetes=Yes **and** BMI$\geq$ 0.4 **and** Prob-Infections$\geq$ 0.2 **then** Diabetes

If Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

Interpretable Decision Sets: Desiderata

- Optimize for the following criteria
  - Recall
  - Precision
  - Distinctness
  - Parsimony
  - Class Coverage

- Recall and Precision □ Accurate predications

- Distinctness, Parsimony, and Class Coverage □Interpretability

IDS: Objective Function

- Parsimony
  - Fewer rules: $f_1(\mathcal{R}) = |\mathcal{S}| - \text{size}(\mathcal{R})$

Number of rules

Number of conditions in the rule

  - Fewer predicates:

$$f_2(\mathcal{R}) = L_{\max} \cdot |\mathcal{S}| - \sum_{r \in \mathcal{R}} \text{length}(r)$$

Maximum no. of conditions in any given Input pattern

Total number of input patterns

IDS: Objective Function

- Distinctness
  - Intra-class overlap:

$$f_3(\mathcal{R}) = N \cdot |S|^2 - \sum_{\substack{r_i, r_j \in \mathcal{R} \\ i \leq j \\ c_i = c_j}} \text{overlap}(r_i, r_j)$$

Total number of data points

Number of points that satisfy both the rules

  - Inter-class overlap:

$$f_4(\mathcal{R}) = N \cdot |S|^2 - \sum_{\substack{r_i, r_j \in \mathcal{R} \\ i \leq j \\ c_i \neq c_j}} \text{overlap}(r_i, r_j)$$

IDS: Objective Function

- Class Coverage

$$f_5(\mathcal{R}) = \sum_{c' \in \mathcal{C}} 1\left(\exists r = (s, c) \in \mathcal{R} \text{ such that } c = c'\right)$$

Check if there exists some rule corresponding to a given class *c*

IDS: Objective Function

- Precision
  - Minimize "incorrect" covers:

$$f_6(\mathcal{R}) = N \cdot |\mathcal{S}| - \sum_{r \in \mathcal{R}} |\text{incorrect-cover}(r)|$$

Given a rule $r = (s,c)$, the no. of data points which satisfy s but do not belong to class c.

IDS: Objective Function

- Recall
  - Encourage at least one "correct" cover per data point:

$$f_7(\mathcal{R}) = \sum_{(\mathbf{x},y)\in\mathcal{D}} 1\left(|\{r|(\mathbf{x},y) \in \text{correct-cover}(r)\}| \geq 1\right)$$

Given a rule $r = (s,c)$, the no. of data points which satisfy s and belong to class c.

IDS: Objective Function

- Complete objective is

$$\underset{\mathcal{R} \subseteq \mathcal{S} \times \mathcal{C}}{\operatorname{argmax}} \sum_{i=1}^{7} \lambda_i f_i(\mathcal{R})$$

- The intra-class and inter-class overlap terms are non-monotone
- The parsimony, overlap, and precision terms are non-normal
- All the component terms are submodular

IDS: Optimization Procedure

- The problem is a non-normal, non-monotone, submodular optimization problem
- Maximizing a non-monotone submodular function is NP-hard
- Local search method which iteratively adds and removes elements until convergence
  - Provides a 2/5 approximation

Inherently Interpretable Models

- Rule Based Models
- Risk Scores
- Generalized Additive Models
- Prototype Based Models
- Attention Based Models

Risk Scores: Motivation

- Risk scores are widely used in medicine and criminal justice
  - E.g., assess risk of mortality in ICU, assess the risk of recidivism
- Adoption decision makers find them easy to understand
- Until very recently, risk scores were constructed manually by domain experts. Can we learn these in a data-driven fashion?

[Ustun and Rudin, 2016]

Risk Scores: Examples

- Recidivism

1.	*Prior Arrests* $\geq$ *2*	1 point		$\cdots$
2.	*Prior Arrests* $\geq$ *5*	1 point	+	$\cdots$
3.	*Prior Arrests for Local Ordinance*	1 point	+	$\cdots$
4.	*Age at Release between 18 to 24*	1 point	+	$\cdots$
5.	*Age at Release* $\geq$ *40*	-1 point	+	$\cdots$
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	$\cdots$

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

- Loan Default

1.	*Call between January and March*	1 point		$\cdots$
2.	*Called Previously*	1 point	+	$\cdots$
3.	*Previous Call was Successful*	1 point	+	$\cdots$
4.	*Employment Indicator* < *5100*	1 point	+	$\cdots$
5.	*3 Month Euribor Rate* $\geq$ *100*	-1 point	+	$\cdots$
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	$\cdots$

SCORE	-1	0	1	2	3	4
RISK	4.7%	11.9%	26.9%	50.0%	73.1%	88.1%

Objective function to learn risk scores

Definition 1 (Risk Score Problem, RiskSlimMINLP)
The risk score problem is a discrete optimization problem with the form:

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & l(\boldsymbol{\lambda}) + C_0 \|\boldsymbol{\lambda}\|_0 \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{L}, \end{aligned} \tag{1}$$

where:

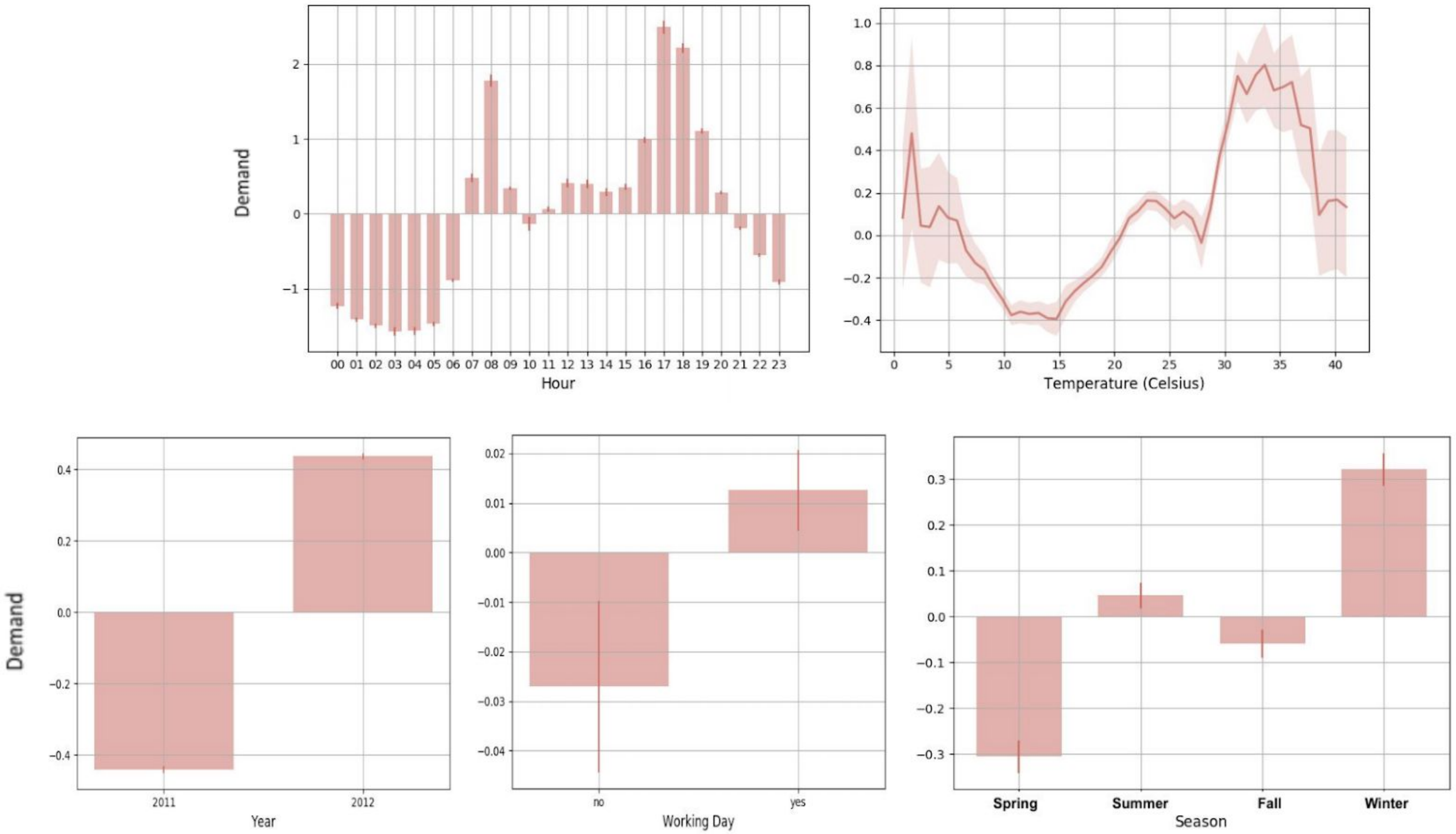
- $l(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{\lambda}, y_i \boldsymbol{x}_i \rangle))$ *is the normalized logistic loss function;*
- $\|\boldsymbol{\lambda}\|_0 = \sum_{j=1}^{d} \mathbb{1}[\lambda_j \neq 0]$ *is the* ℓ_0*-seminorm;*
- $\mathcal{L} \subset \mathbb{Z}^{d+1}$ *is a set of feasible coefficient vectors (user-provided);*
- $C_0 > 0$ *is a trade-off parameter to balance fit and sparsity (user-provided).*

Above turns out to be a mixed integer program, and is optimized using a cutting plane method and a branch-and-bound technique.

Inherently Interpretable Models

- Rule Based Models
- Risk Scores
- Generalized Additive Models
- Prototype Based Models
- Attention Based Models

Generalized Additive Models (GAMs)



Formulation and Characteristics of GAMs

Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + \ldots + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + \ldots + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, \ldots, x_n)$	+	+++

g is a link function; E.g., identity function in case of regression;
log (y/1 – y) in case of classification;

f_i is a shape function

GAMs and GA^2Ms

- While GAMs model first order terms, GA^2Ms model second order feature interactions as well.

$$\text{GAMs:}\quad g(y) = \beta_0 + \sum f_j(x_j)$$

$$\text{GA}^2\text{Ms:}\quad g(y) = \beta_0 + \sum f_j(x_j) + \sum_{i \neq j} f_{i,j}(x_i, x_j)$$

GAMs and GA^{2}Ms

- Learning:
  - Represent each component as a spline
  - Least squares formulation; Optimization problem to balance smoothness and empirical error

- GA^{2}Ms: Build GAM first and and then detect and rank all possible pairs of interactions in the residual
  - Choose top k pairs
  - k determined by CV

Inherently Interpretable Models

- Rule Based Models
- Risk Scores
- Generalized Additive Models
- Prototype Based Models
- Attention Based Models

[Bien et. al., 2012]

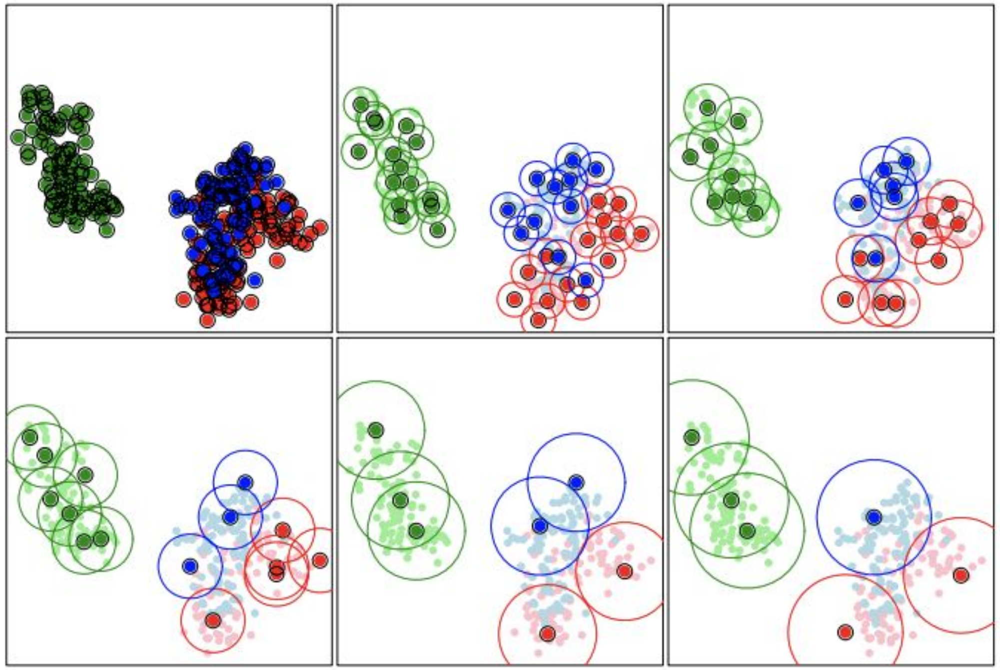
Prototype Selection for Interpretable Classification

- The goal here is to identify K prototypes (instances) from the data s.t. a new instance which will be assigned the same label as the closest prototype will be correctly classified (with a high probability)

- Let each instance "cover" the ϵ - neighborhood around it.

- Once we define the neighborhood covered by each instance, this problem becomes similar to the problem of finding rule sets, and can be solved analogously.

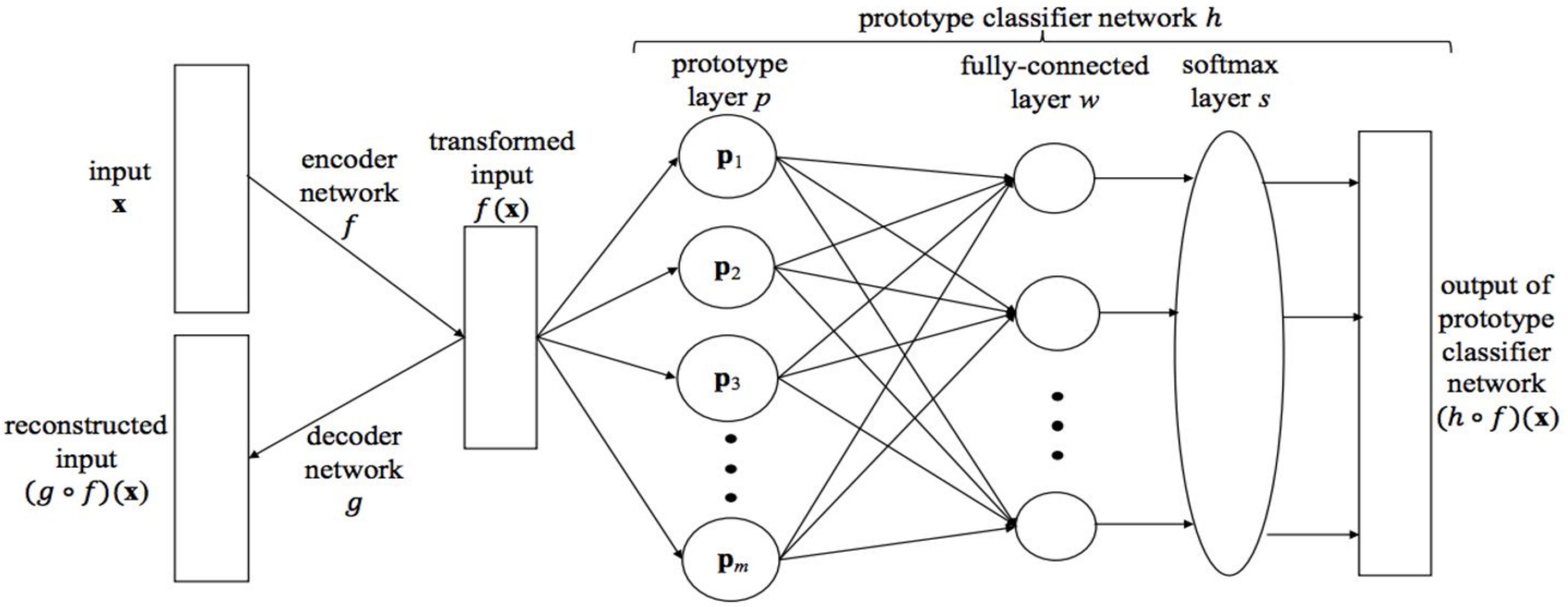
Prototype Selection for Interpretable Classification



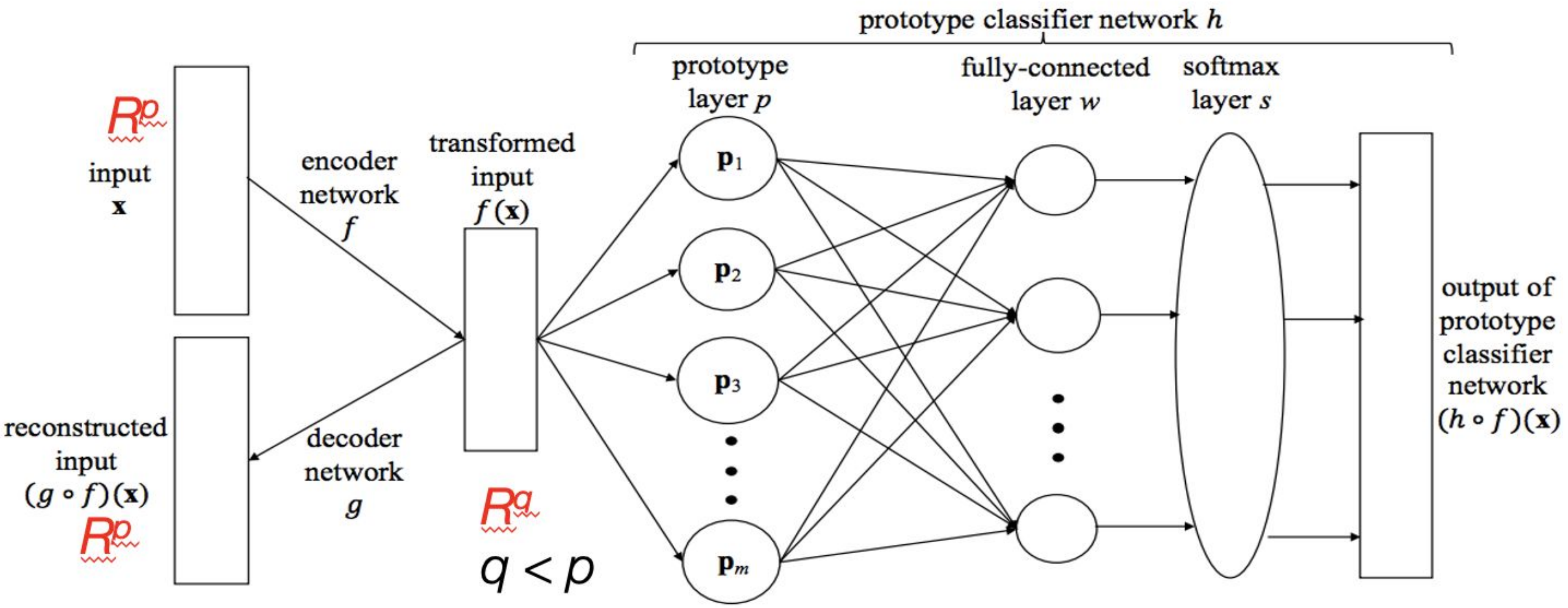
Given a value for ε, the choice of $\mathcal{P}_1, \ldots, \mathcal{P}_L$ induces L partial covers of the training points by ε-balls. Here ε is varied from the smallest (top-left panel) to approximately the median interpoint distance (bottom-right panel).

[Li et. al. 2017, Chen et. al. 2019]

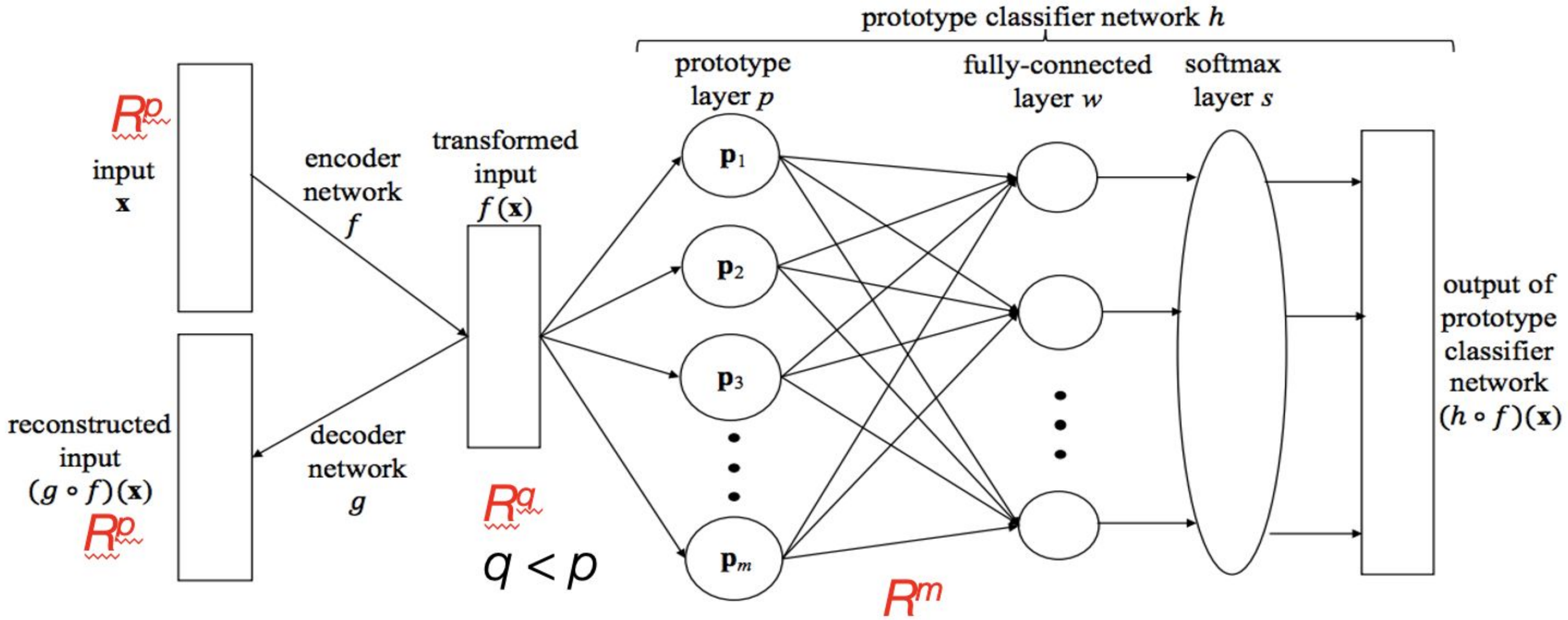
Prototype Layers in Deep Learning Models



Prototype Layers in Deep Learning Models



Prototype Layers in Deep Learning Models

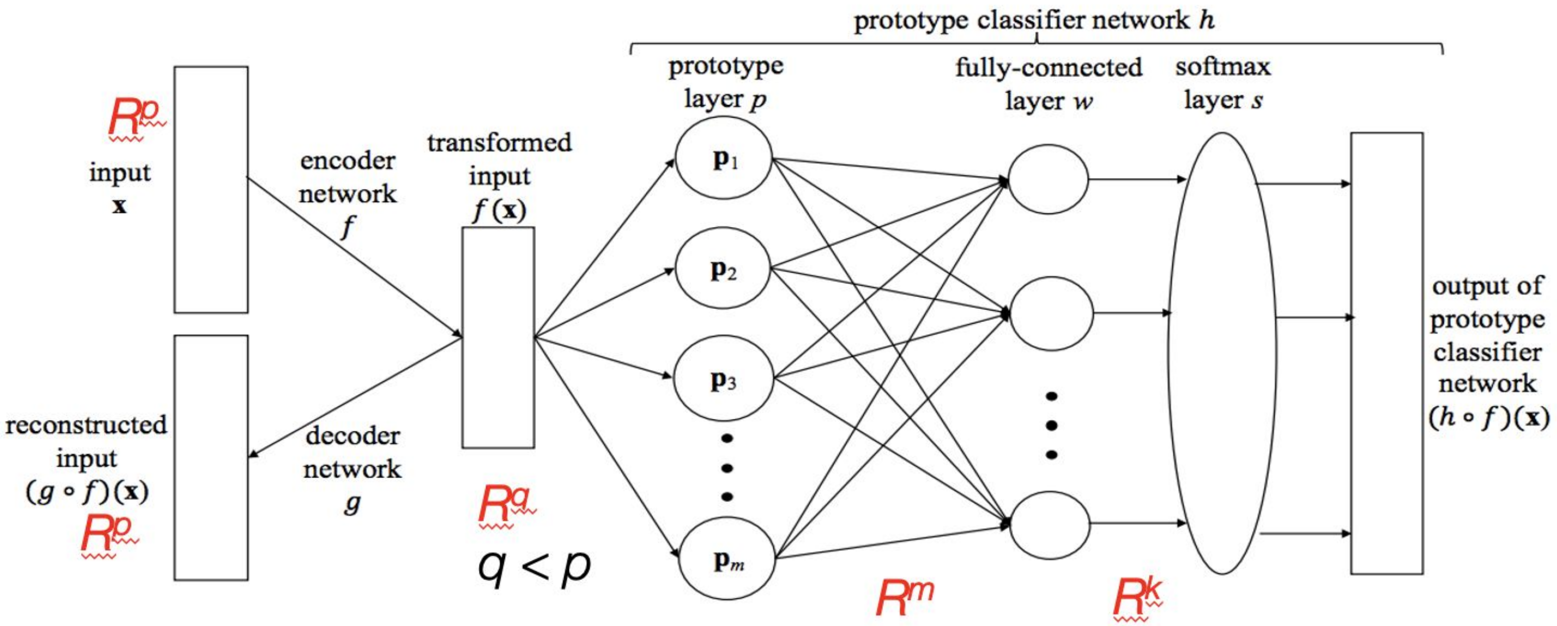


Prototype layer is responsible for computing the prototypes

$$\mathbf{z} = f(\mathbf{x}) \quad p(\mathbf{z}) = \left[\|\mathbf{z}-\mathbf{p}_1\|_2^2, \quad \|\mathbf{z}-\mathbf{p}_2\|_2^2, \quad \dots \quad \|\mathbf{z}-\mathbf{p}_m\|_2^2\right]^\top$$

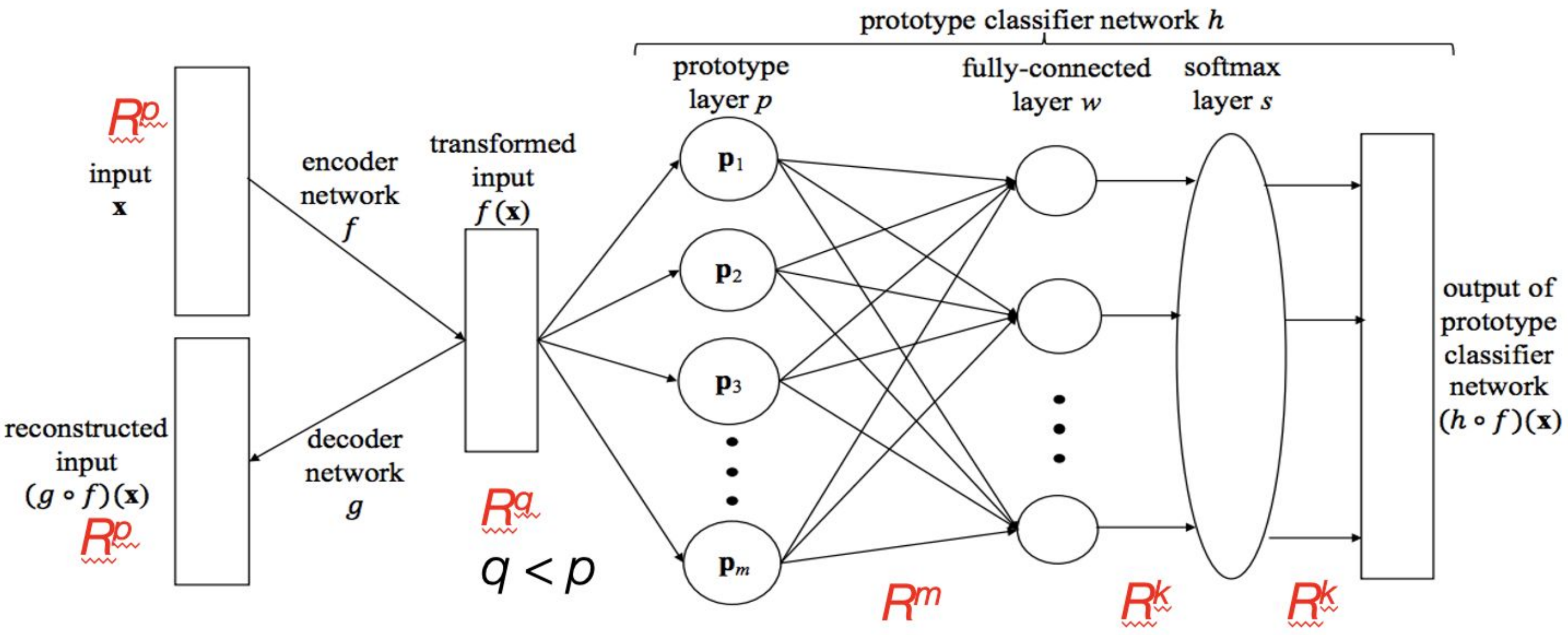
Each node in layer p computes one of the above elements

Prototype Layers in Deep Learning Models



- The fully connected layer computes weighted sums of the distances $\|\mathbf{z} - \mathbf{p}_j\|_2^2$: $Wp(\mathbf{z})$
- W is a k x m matrix

Prototype Layers in Deep Learning Models



The weighted sums $Wp(\mathbf{z})$ are normalized by the softmax layer to output a probability distribution over K classes

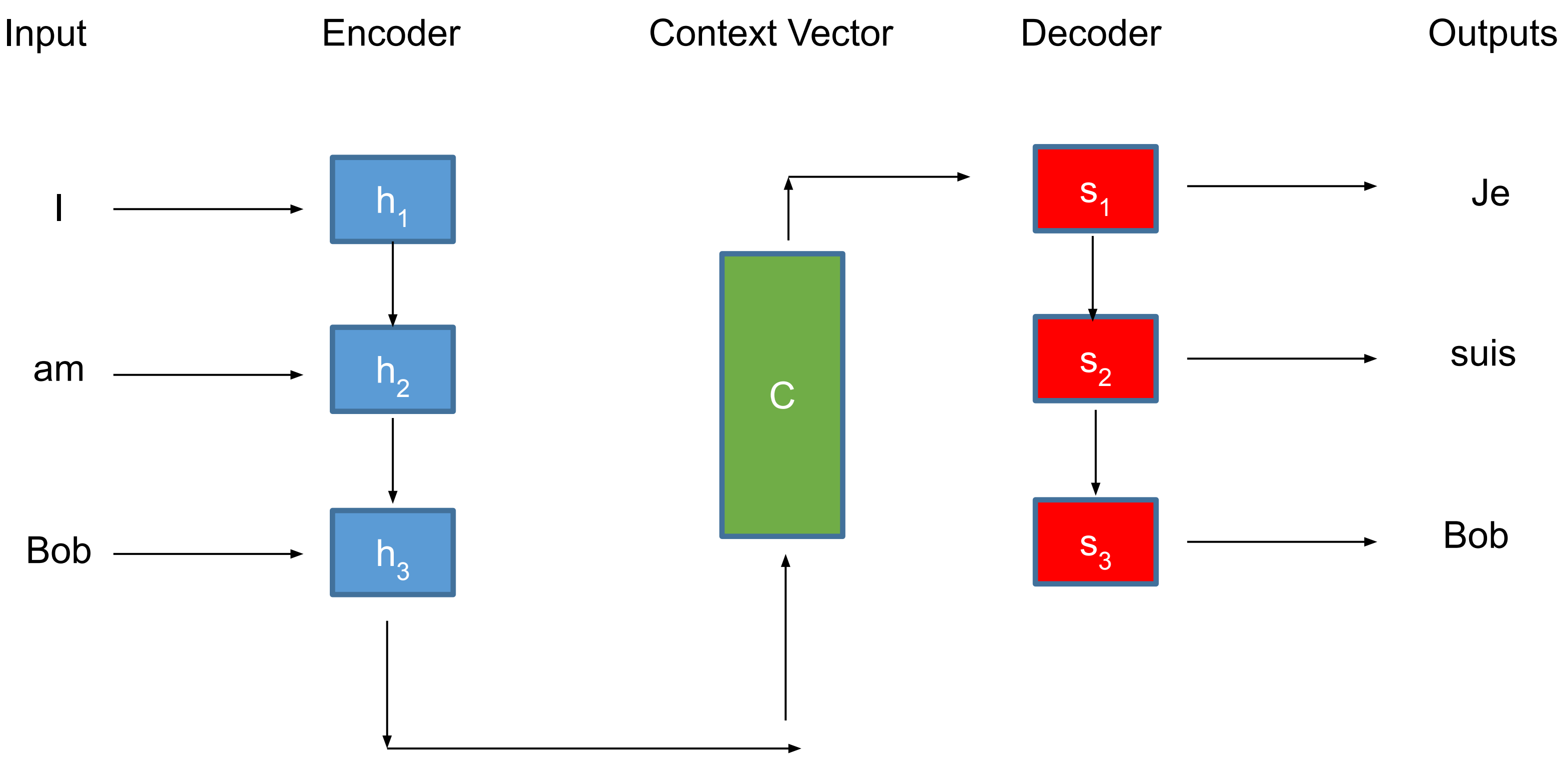
Inherently Interpretable Models

- Rule Based Models
- Risk Scores
- Generalized Additive Models
- Prototype Based Models
- Attention Based Models

[Bahdanau et. al. 2016; Xu et. al. 2015]

Attention Layers in Deep Learning Models

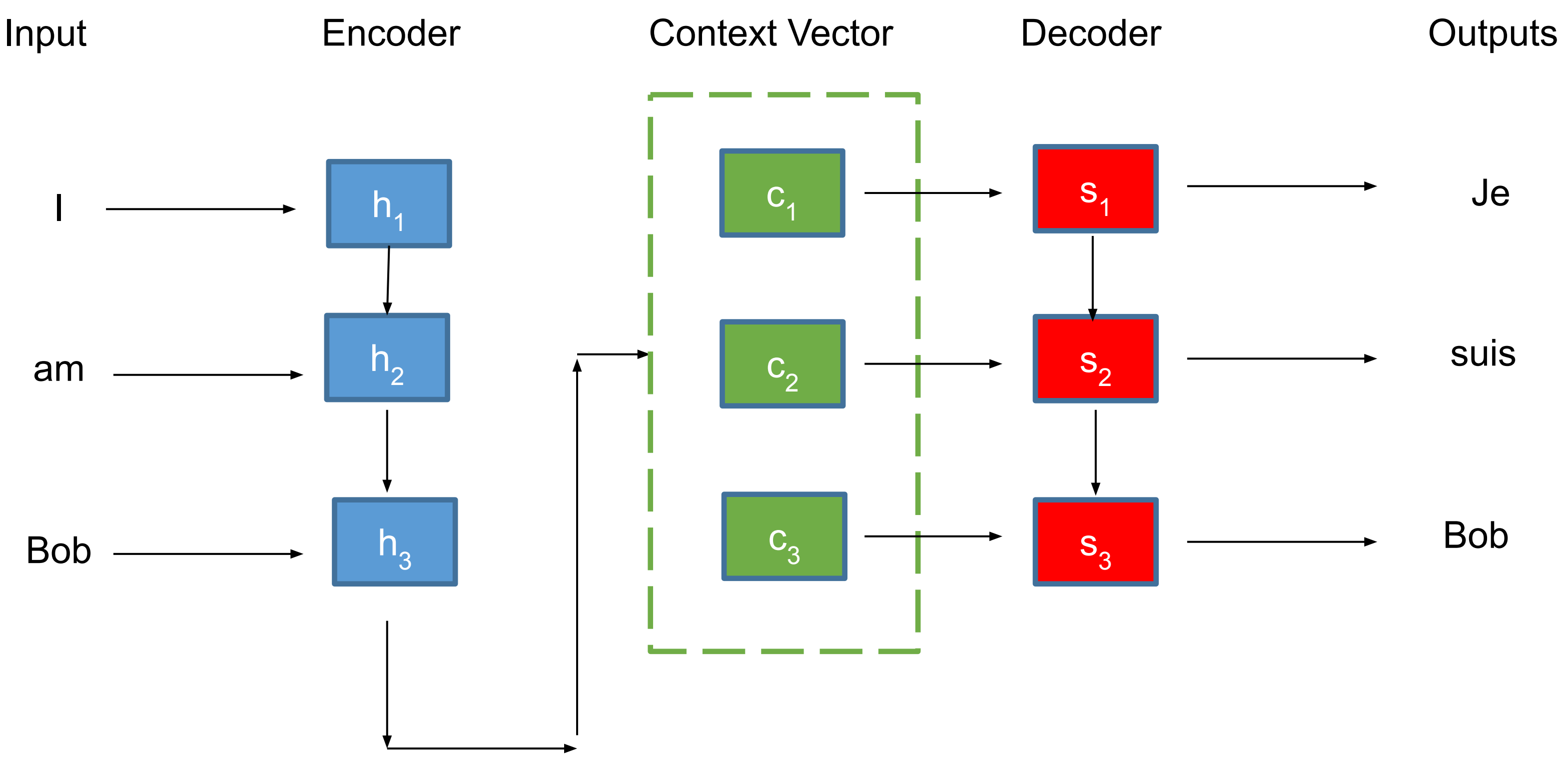
- Let us consider the example of machine translation



[Bahdanau et. al. 2016; Xu et. al. 2015]

Attention Layers in Deep Learning Models

- Let us consider the example of machine translation



Attention Layers in Deep Learning Models

- Context vector corresponding to s_i can be written as follows:

$$c_i = \sum_{j=1}^{3} a_{ij} h_j$$

a_{ij}

- captures the attention placed on input token j when determining the decoder hidden state s_i; it can be computed as a softmax of the "match" between s_{i-1} and h_j

Inherently Interpretable Models

- Rule Based Models
- Risk Scores
- Generalized Additive Models
- Prototype Based Models
- Attention Based Models

Agenda

- ❑ Inherently Interpretable Models
- ❑ Post hoc Explanation Methods
- ❑ Evaluating Model Interpretations/Explanations
- ❑ Empirically & Theoretically Analyzing Interpretations/Explanations
- ❑ Future of Model Understanding

What is an Explanation?

What is an Explanation?

Definition: Interpretable description of the model behavior

Classifier

Faithful

Explanation

Understandable

User

[Lipton 2016]

What is an Explanation?

Definition: Interpretable description of the model behavior

Classifier

Send all the model parameters θ?

Send many example predictions?

Summarize with a program/rule/tree

Select most important features/points

Describe how to *flip* the model prediction

...

User

Local Explanations vs. Global Explanations

Local Explanations	Global Explanations
Explain individual predictions	Explain complete behavior of the model
Help unearth biases in the *local neighborhood* of a given instance	Help shed light on *big picture biases* affecting larger subgroups
Help vet if individual predictions are being made for the right reasons	Help vet if the model, at a high level, is suitable for deployment

Approaches for Post hoc Explainability

Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Approaches for Post hoc Explainability

Local Explanations

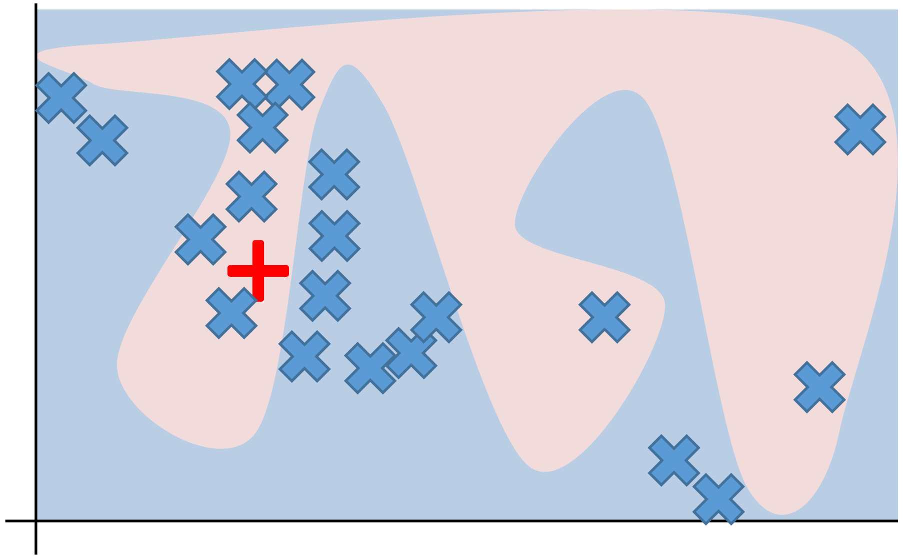
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

LIME: Local Interpretable Model-Agnostic Explanations

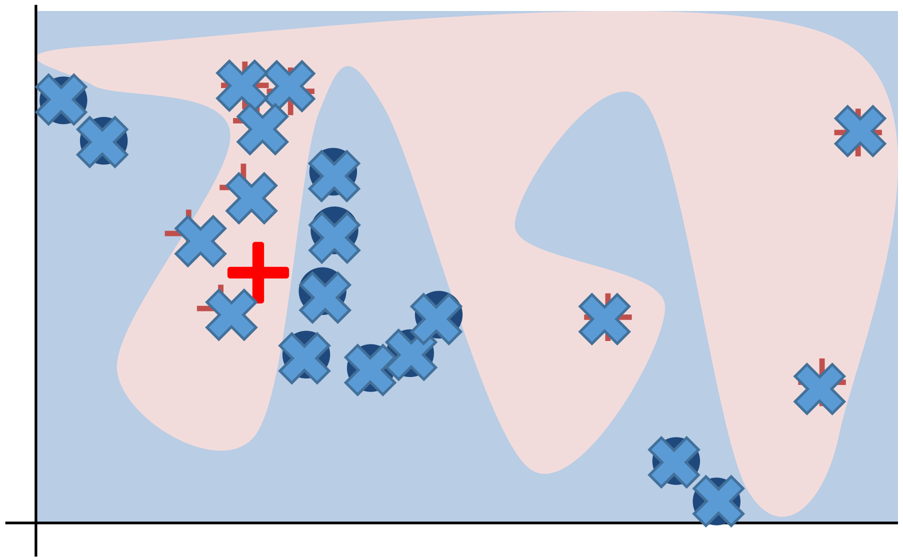
1. Sample points around x_i



[Ribeiro et al. 2016]

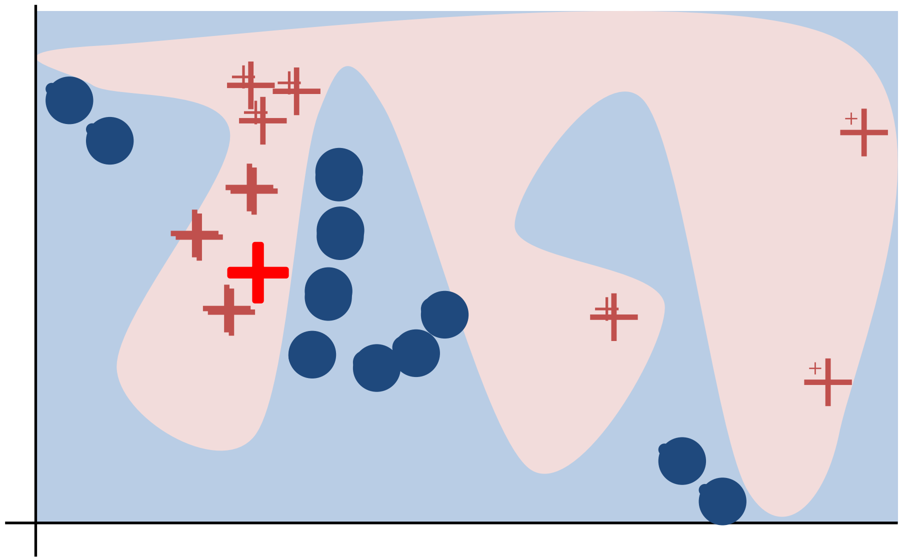
LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample



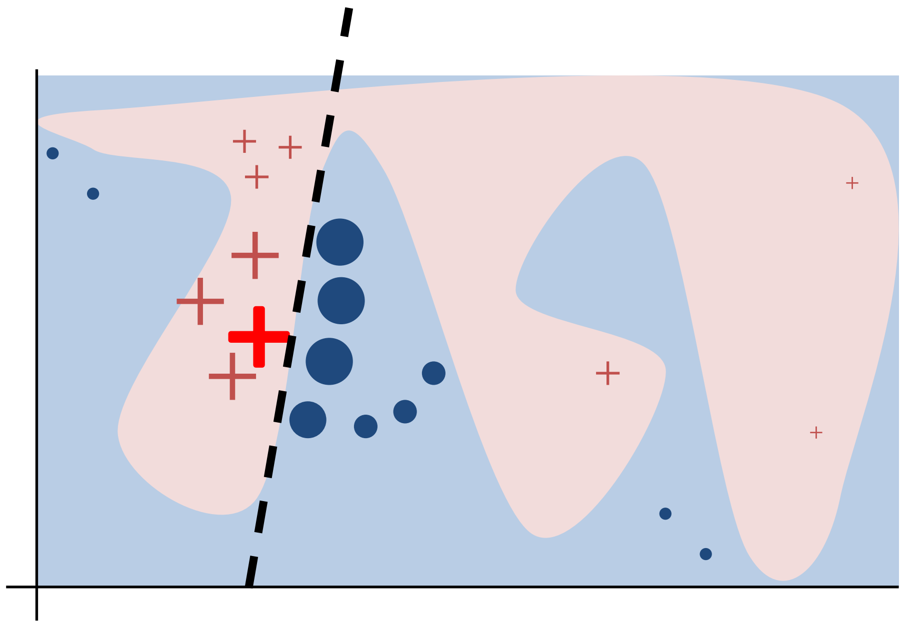
LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i



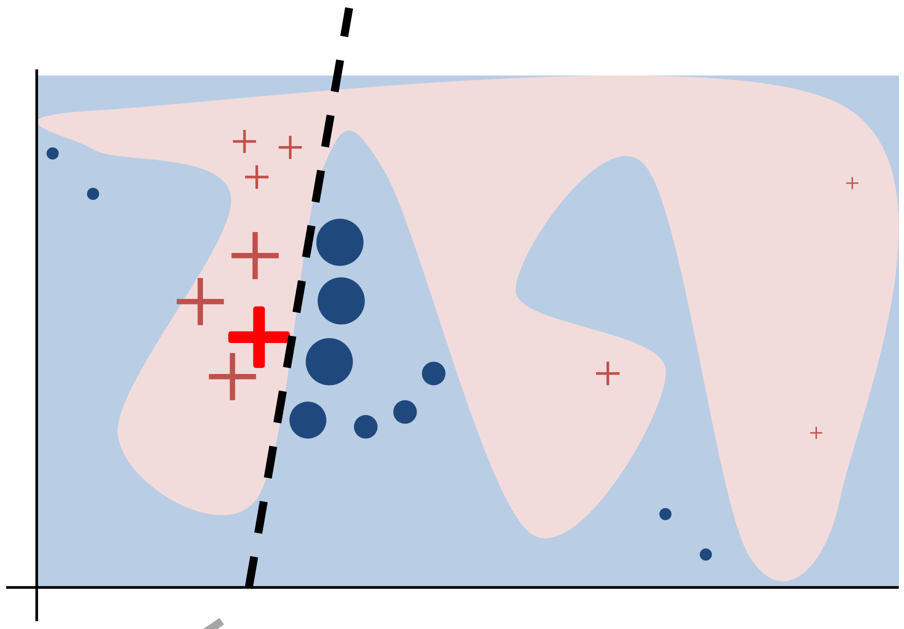
LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples



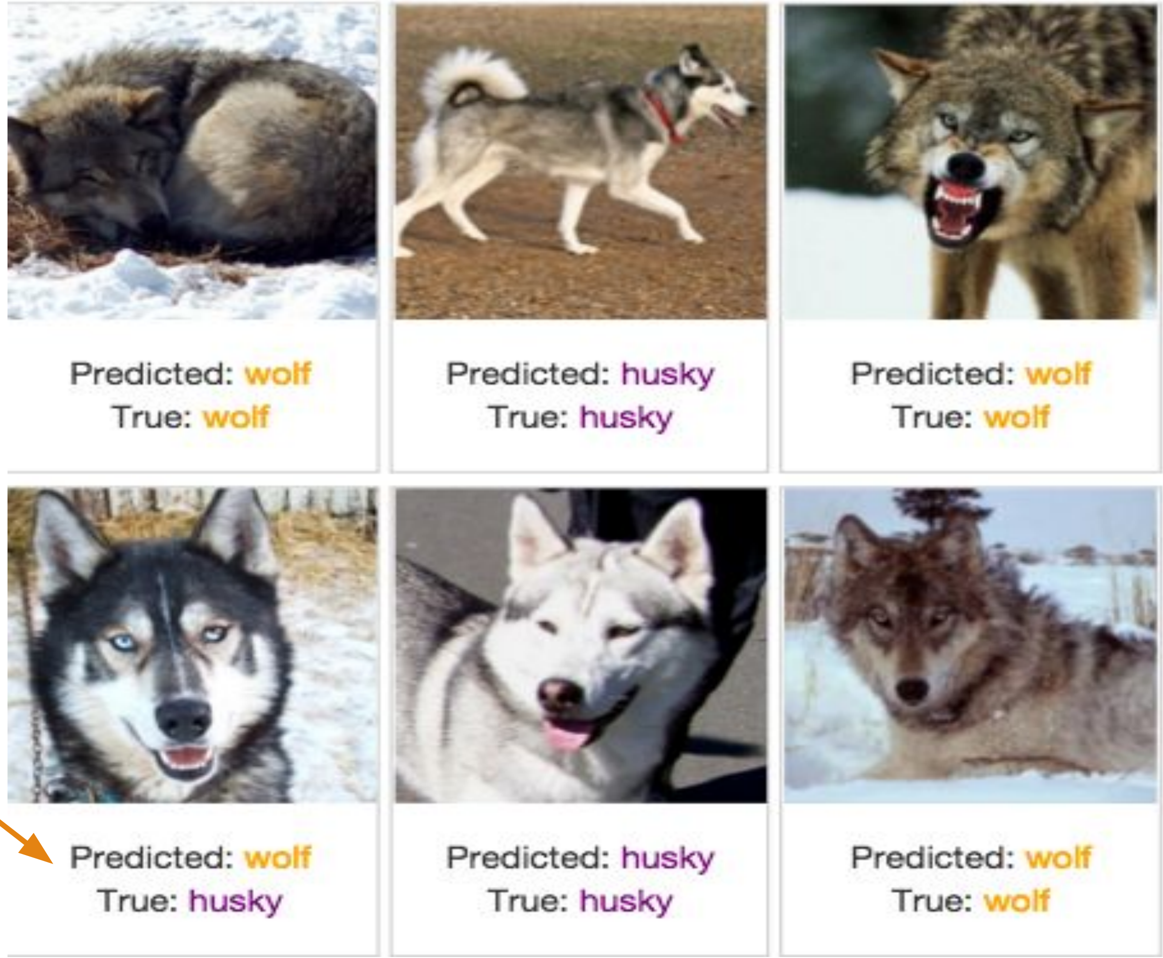
LIME: Local Interpretable Model-Agnostic Explanations

1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples
5. Use simple linear model to explain

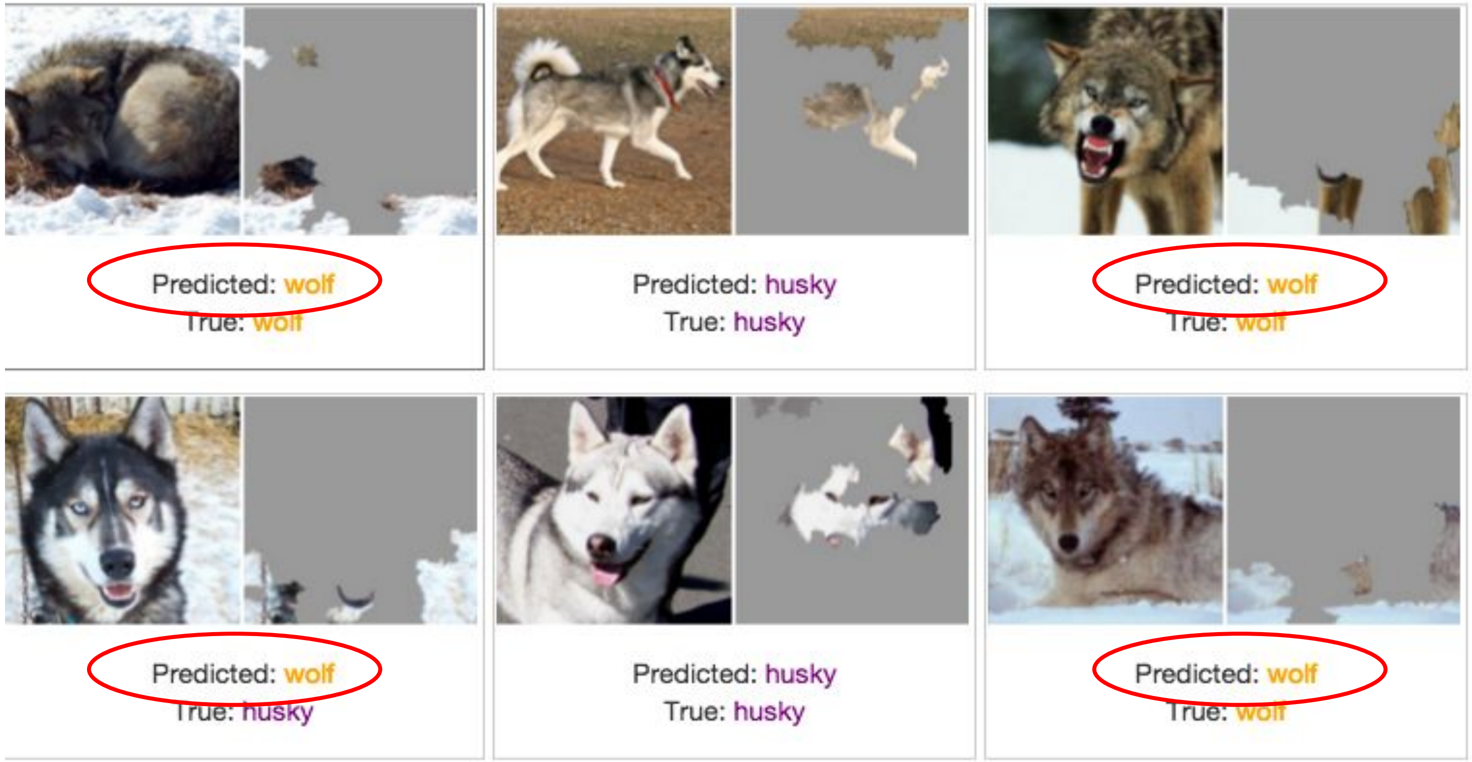


Predict Wolf vs Husky

Only 1 mistake!



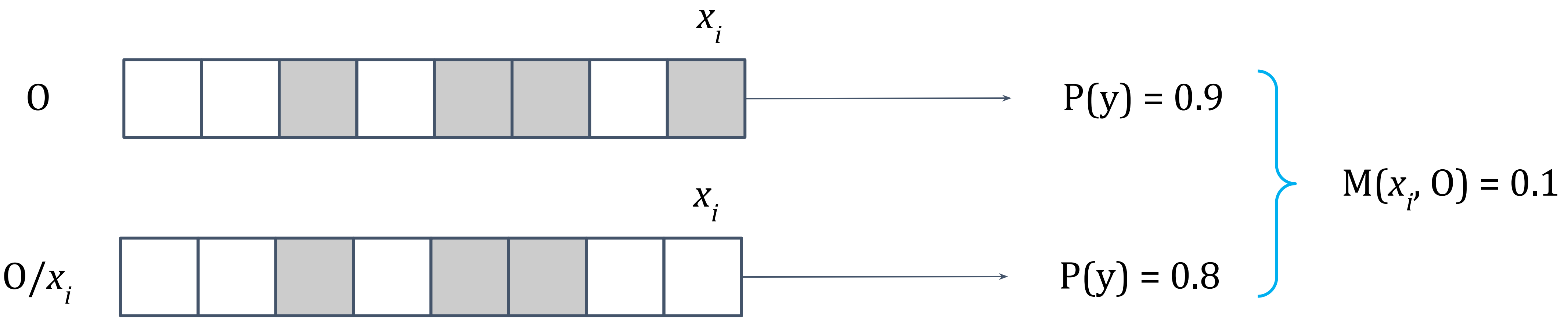
Predict Wolf vs Husky



We’ve built a great snow detector…

SHAP: Shapley Values as Importance

Marginal contribution of each feature towards the prediction, averaged over all possible permutations.



Attributes the prediction to each of the features.

Approaches for Post hoc Explainability

Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

[Ribeiro et al. 2018]

Anchors

- Perturb a given instance x to generate local neighborhood

- Identify an “anchor” rule which has the maximum coverage of the local neighborhood and also achieves a high precision.

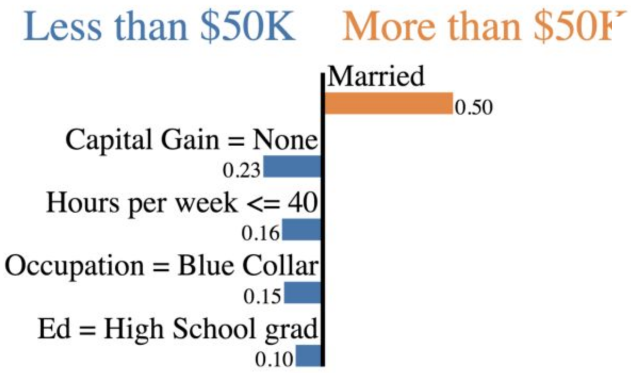
[Ribeiro et al. 2018]

Salary Prediction

$28 < \text{Age} \leq 37$
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week ≤ 40.00
Country = United-States

$P(\textbf{Salary} > \$50\textbf{K}) = 0.57$

(a) Instance and prediction



(b) LIME explanation

IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** $28 < \text{Age} \leq 37$
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > $50K

(c) An *anchor* explanation

Approaches for Post hoc Explainability

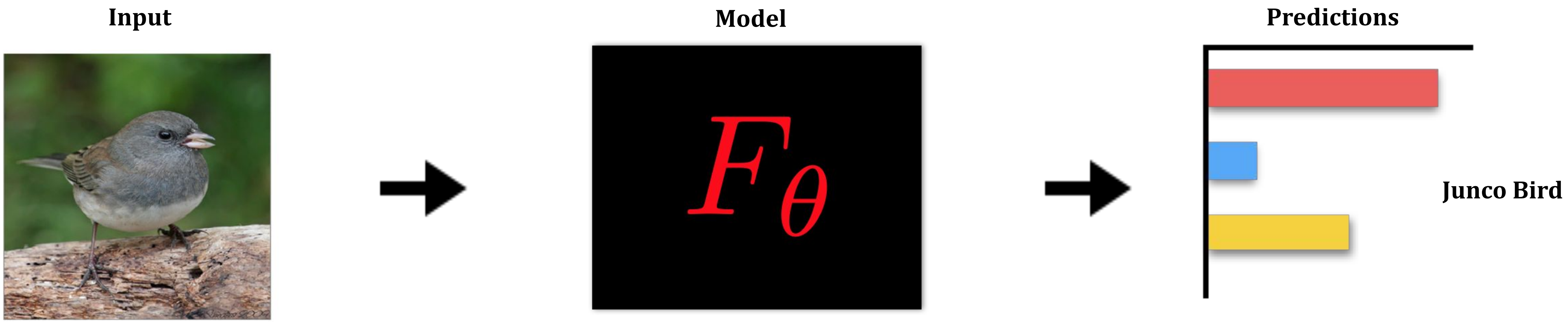
Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

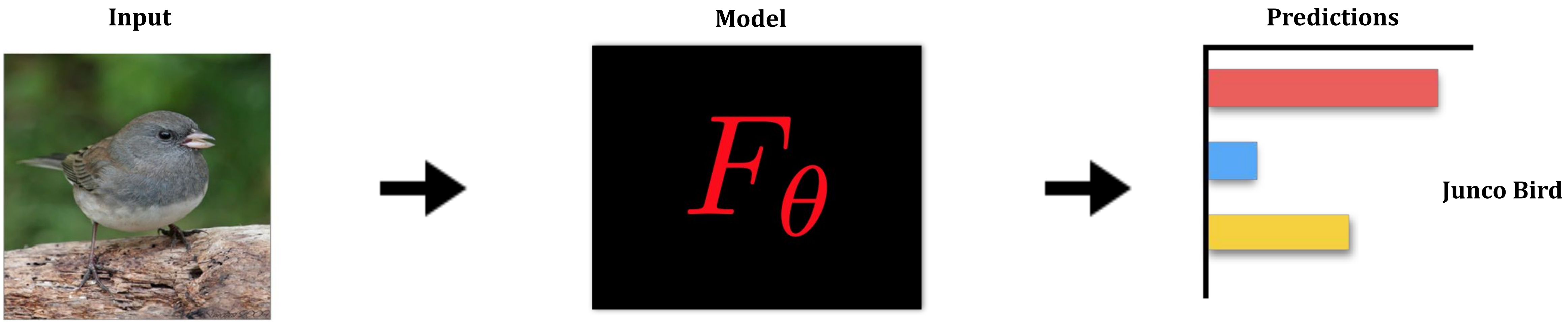
Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Saliency Map Overview



Saliency Map Overview



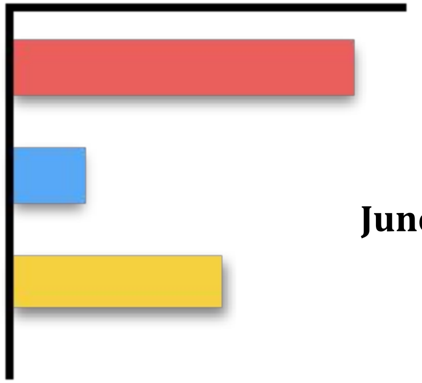
What parts of the input are most relevant for the model's prediction: 'Junco Bird'?

Saliency Map Overview

Input

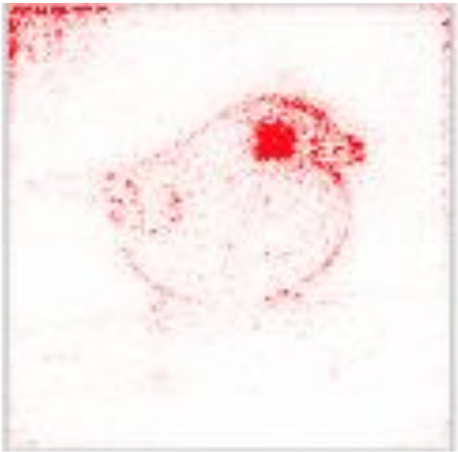
Model

Predictions



Junco Bird

What parts of the input are most relevant for the model's prediction: ‘Junco Bird’?



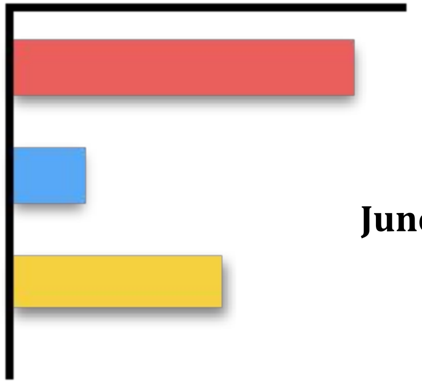
Modern DNN Setting

Input

Model

F_θ

Predictions



Junco Bird

$$F : \mathbb{R}^d \rightarrow \mathbb{R}^c$$ **Model**

$$F_i : \mathbb{R}^d \rightarrow \mathbb{R}$$ **class specific logit**

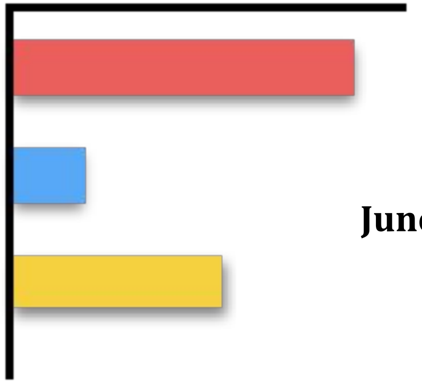
Input-Gradient

Input

Model

$$F_\theta$$

Predictions



Junco Bird

Input-Gradient

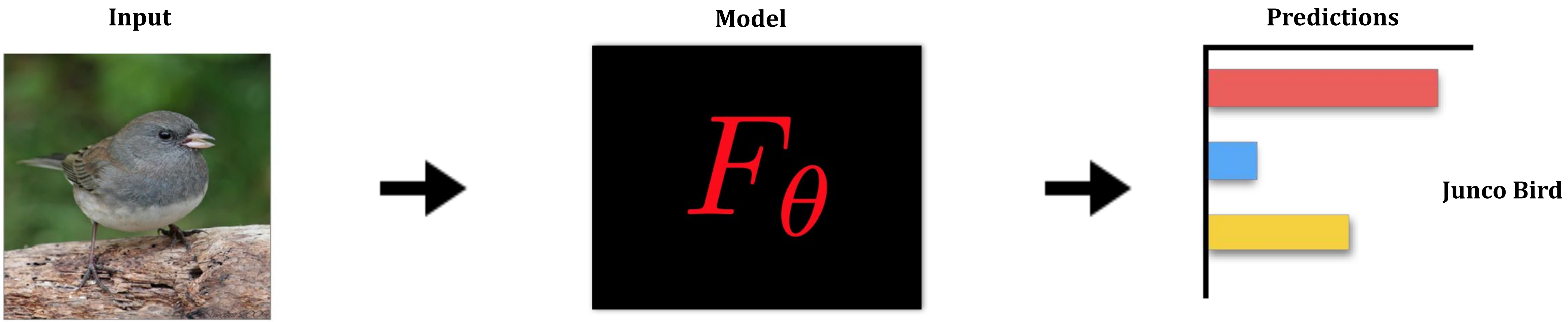
$$\nabla_x F_i(x) \rightarrow \in \mathbb{R}^d$$

Input **Logit**

Same dimension as the input.

[Baehrens et. al. 2010](); [Simonyan et. al. 2014]() .

Input-Gradient



Input-Gradient

$$\nabla_x F_i(x)$$

Logit

Input

Visualize as a heatmap

[Baehrens et. al. 2010](); [Simonyan et. al. 2014]() .

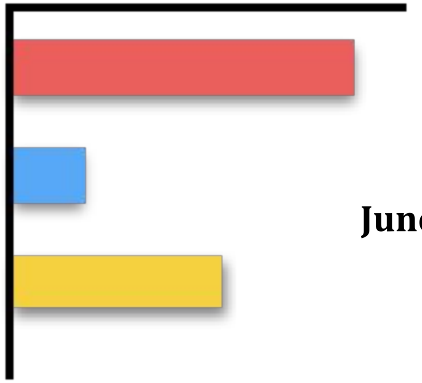
Input-Gradient

Input

Model

$$F_\theta$$

Predictions



Junco Bird

Input-Gradient

$$\nabla_x F_i(x)$$

Logit

Input

Challenges

- **Visually noisy & difficult to interpret.**
- **'Gradient saturation.'**

[Shrikumar et. al. 2017](#).

[Baehrens et. al. 2010](#); [Simonyan et. al. 2014](#) .

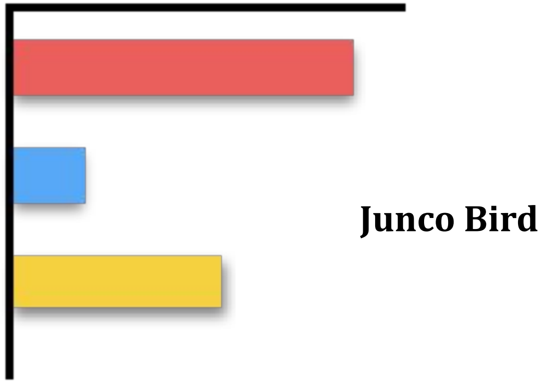
SmoothGrad

Input

Model

$$F_\theta$$

Predictions



SmoothGrad

$$\frac{1}{N}\sum_{i}^{N}\nabla_{(x+\epsilon)}F_i(x+\epsilon)$$

Gaussian noise

Average Input-gradient of 'noisy' inputs.

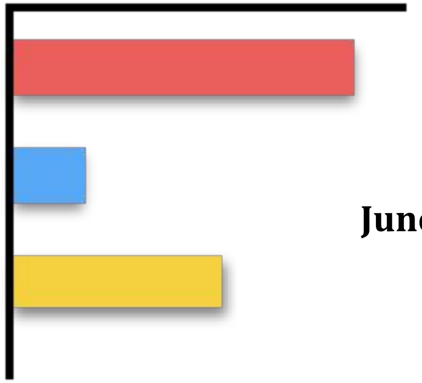
[Smilkov et. al. 2017]()

SmoothGrad

Input

Model

Predictions



Junco Bird

Average Input-gradient of ‘noisy’ inputs.

SmoothGrad

$$\frac{1}{N}\sum_{i}^{N}\nabla_{(x+\epsilon)}F_i(x+\epsilon)$$

Gaussian noise

[Smilkov et. al. 2017]()

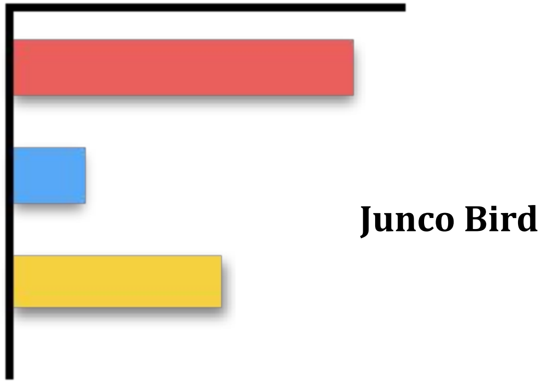
Integrated Gradients

Input

Model

$$F_\theta$$

Predictions



$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

Baseline input

Path integral: 'sum' of interpolated gradients

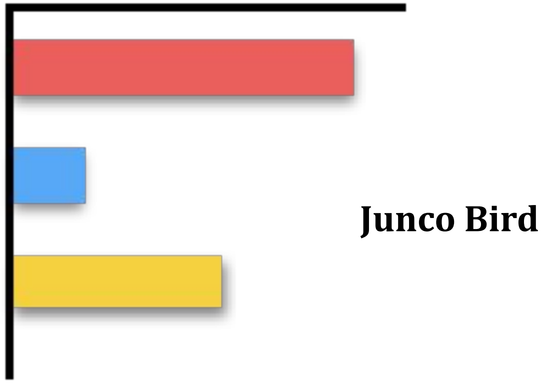
Sundararajan et. al. 2017

Integrated Gradients

Input

Model

Predictions



$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

Baseline input

Path integral: ‘sum’ of interpolated gradients

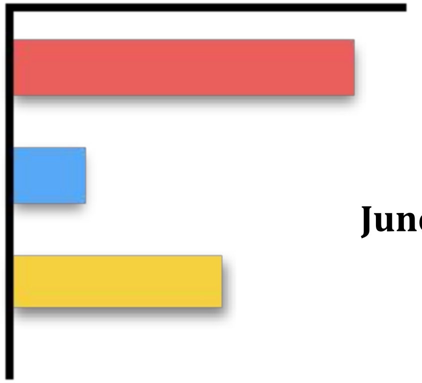
[Sundararajan et. al. 2017]

Gradient-Input

Input

Model

Predictions



Gradient-Input

$$\nabla_x F(x) \odot x$$

Input gradient

Input

Element-wise product of input-gradient and input.

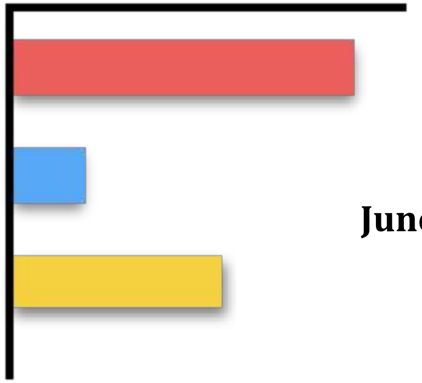
Shrikumar et. al. 2017, Ancona et. al. 2018.

Gradient-Input

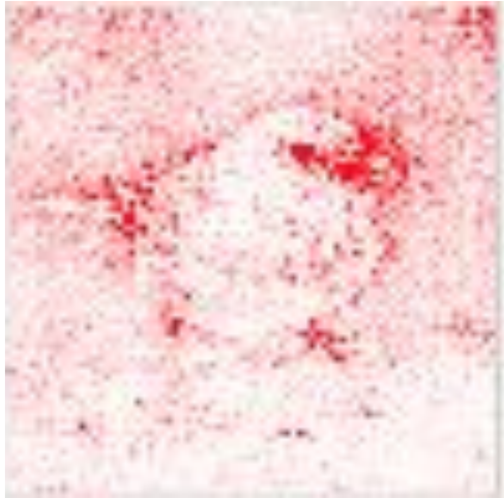
Input

Model

Predictions



Junco Bird



Gradient-Input

$$\nabla_x F(x) \odot x$$

logit gradient

Input

Element-wise product of input-gradient and input.

[Shrikumar et. al. 2017](), [Ancona et. al. 2018.]()

Approaches for Post hoc Explainability

Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

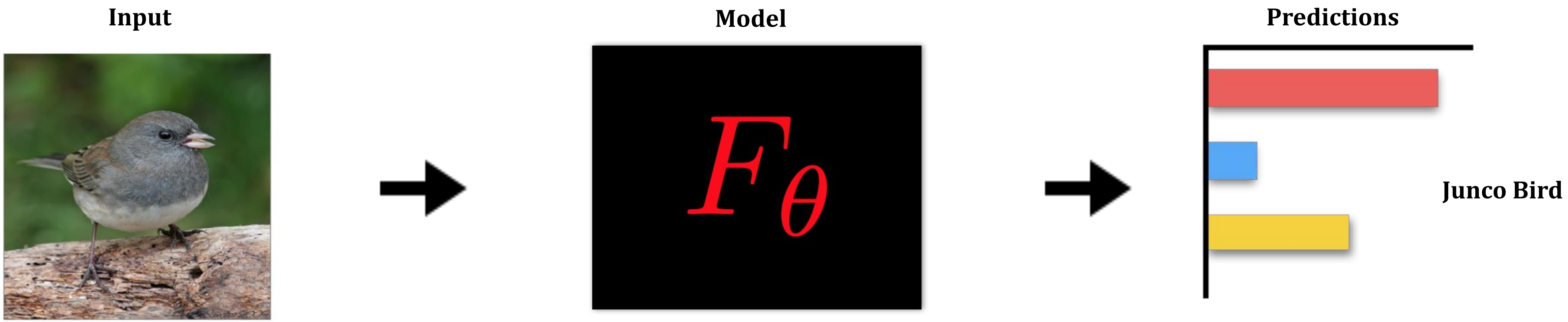
Prototypes/Example Based Post hoc Explanations

Use examples (synthetic or natural) to explain individual predictions

- Influence Functions (**Koh & Liang 2017**)
  - Identify instances in the training set that are responsible for the prediction of a given test instance

- Activation Maximization (**Erhan et al. 2009**)
  - Identify examples (synthetic or natural) that strongly activate a function (neuron) of interest

Training Point Ranking via Influence Functions



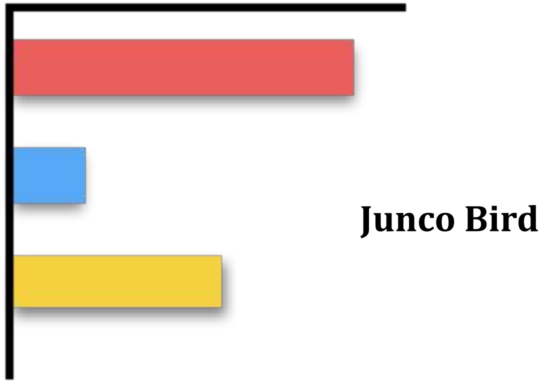
Which training data points have the most ‘*influence*’ on the test loss?

Training Point Ranking via Influence Functions

Input

Model

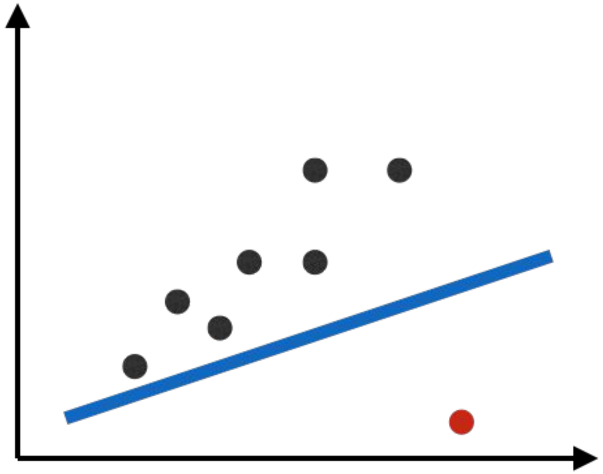
Predictions

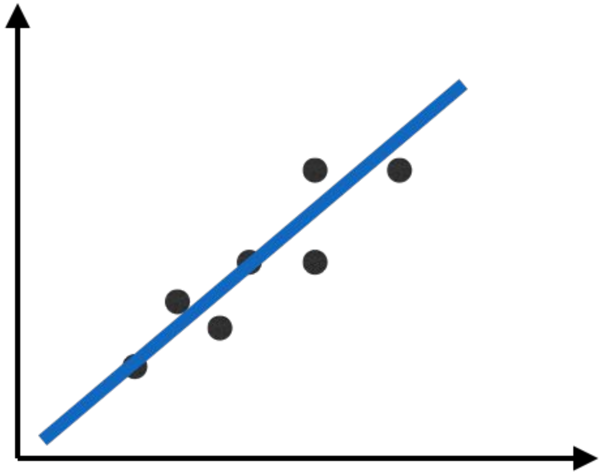


Which training data points have the most '*influence*' on the test loss?

Training Point Ranking via Influence Functions

Influence Function: classic tool used in robust statistics for assessing the effect of a sample on regression parameters ([Cook & Weisberg, 1980](#)).





Instead of refitting model for every data point, **Cook's distance** provides analytical alternative.

Training Point Ranking via Influence Functions

Koh & Liang (2017) extend the 'Cook's distance' insight to modern machine learning setting.

$z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ $z_j = (x_j, y_j) \leftarrow$ **Training sample point** z_{test}

Koh & Liang 2017

Training Point Ranking via Influence Functions

Koh & Liang (2017) extend the 'Cook's distance' insight to modern machine learning setting.

$$z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \qquad z_j = (x_j, y_j) \leftarrow \text{Training sample point} \qquad z_{\text{test}}$$

ERM Solution

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta)$$

UpWeighted ERM Solution

$$\hat{\theta}_{\epsilon, z_j} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta) + \epsilon \ell(z_j; \theta) \qquad \epsilon = -\frac{1}{n}$$

Training Point Ranking via Influence Functions

Koh & Liang (2017) extend the 'Cook's distance' insight to modern machine learning setting.

$$z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \qquad z_j = (x_j, y_j) \qquad z_{\text{test}}$$

Training sample point

ERM Solution

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta)$$

UpWeighted ERM Solution

$$\hat{\theta}_{\epsilon, z_j} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta) + \epsilon \ell(z_j; \theta) \qquad \epsilon = -\frac{1}{n}$$

Influence of Training Point on Parameters

$$\mathcal{I}_{z_j} = \left. \frac{d\hat{\theta}_{\epsilon, z_j}}{d\epsilon} \right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta \ell(z_j, \hat{\theta})$$

Influence of Training Point on Test-Input's loss

$$\mathcal{I}_{z_j, z_{\text{test}}, \text{loss}} = -\nabla_\theta \ell(z_{\text{test}}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta \ell(z_j, \hat{\theta})$$

Koh & Liang 2017

Training Point Ranking via **Influence Functions**

Applications:

- compute self-influence to identify mislabelled examples;
- diagnose possible domain mismatch;
- craft training-time poisoning examples.

[Koh & Liang 2017]

Challenges and Other Approaches

Influence function Challenges:

1. **scalability**: computing hessian-vector products can be tedious in practice.

2. **non-convexity:** possibly loose approximation for deeper networks (Basu et. al. 2020).

Challenges and Other Approaches

Influence function Challenges:

1. **scalability**: computing hessian-vector products can be tedious in practice.

2. **non-convexity:** possibly loose approximation for 'deeper' networks (Basu et. al. 2020).

Alternatives:

- **Representer Points** (Yeh et. al. 2018).

- **TracIn** (Pruthi et. al. appearing at NeuRIPs 2020).

Activation Maximization

These approaches identify examples, synthetic or natural, that **strongly activate a function (neuron) of interest**.

Activation Maximization

These approaches identify examples, synthetic or natural, that **strongly activate a function (neuron) of interest**.

Implementation Flavors:

- Search for **natural examples within a specified set** (training or validation corpus) that strongly activate a neuron of interest;
- **Synthesize examples**, typically via gradient descent, that strongly activate a neuron of interest.

Feature Visualization

Dataset Examples show us what neurons respond to in practice

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

Baseball—or stripes?	Animal faces—or snouts?	Clouds—or fluffiness?	Buildings—or sky?
mixed4a, Unit 6	*mixed4a, Unit 240*	*mixed4a, Unit 453*	*mixed4a, Unit 492*

[Olah et. al. 2017]

Approaches for Post hoc Explainability

Local Explanations

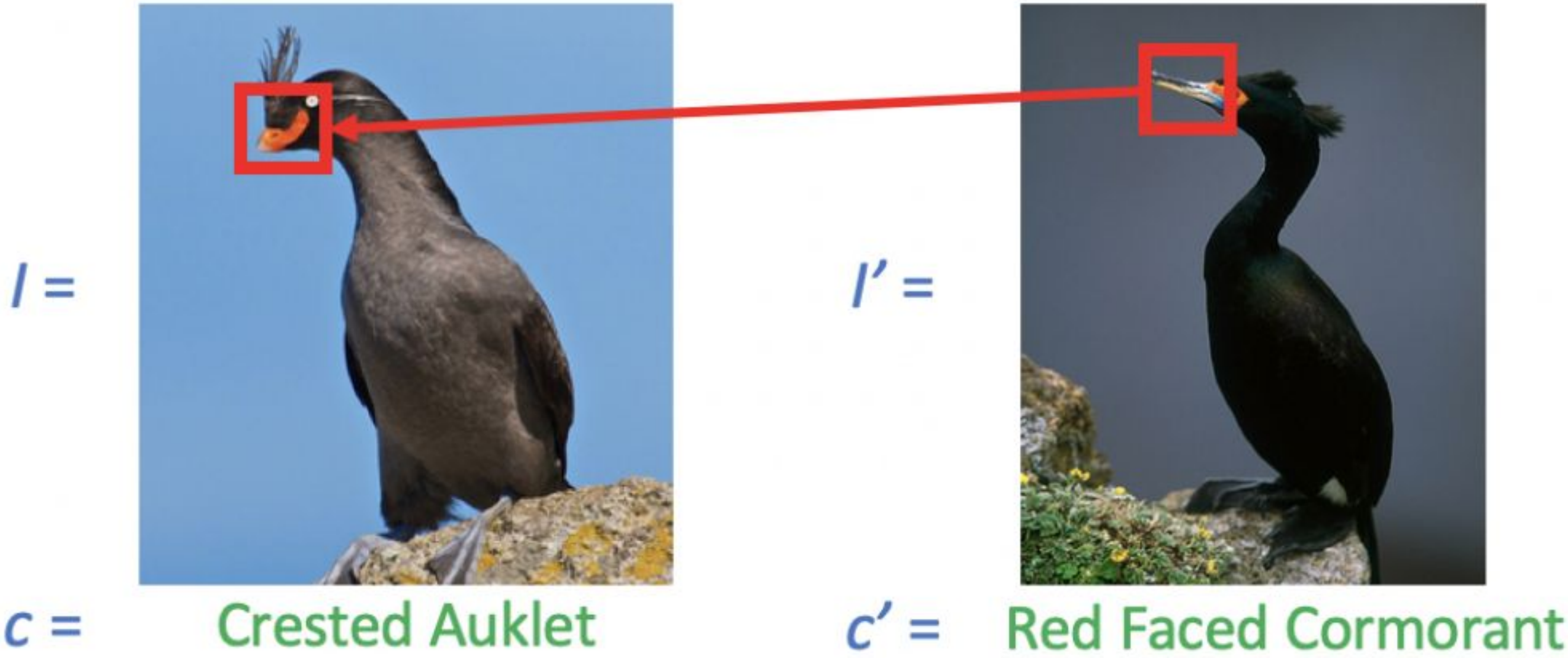
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Counterfactual Explanations

What features need to be changed and by how much to flip a model's prediction?



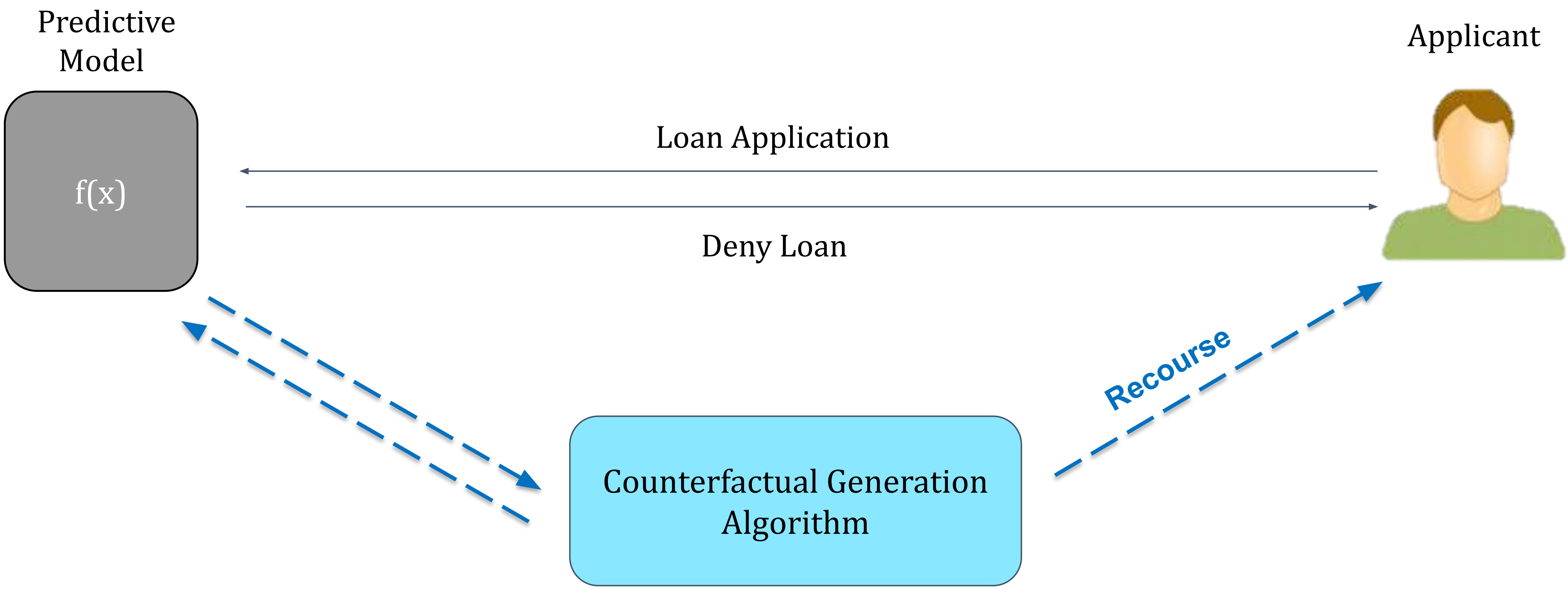
[Goyal et. al., 2019]

Counterfactual Explanations

As ML models increasingly deployed to make high-stakes decisions (e.g., loan applications), it becomes important to provide recourse to affected individuals.

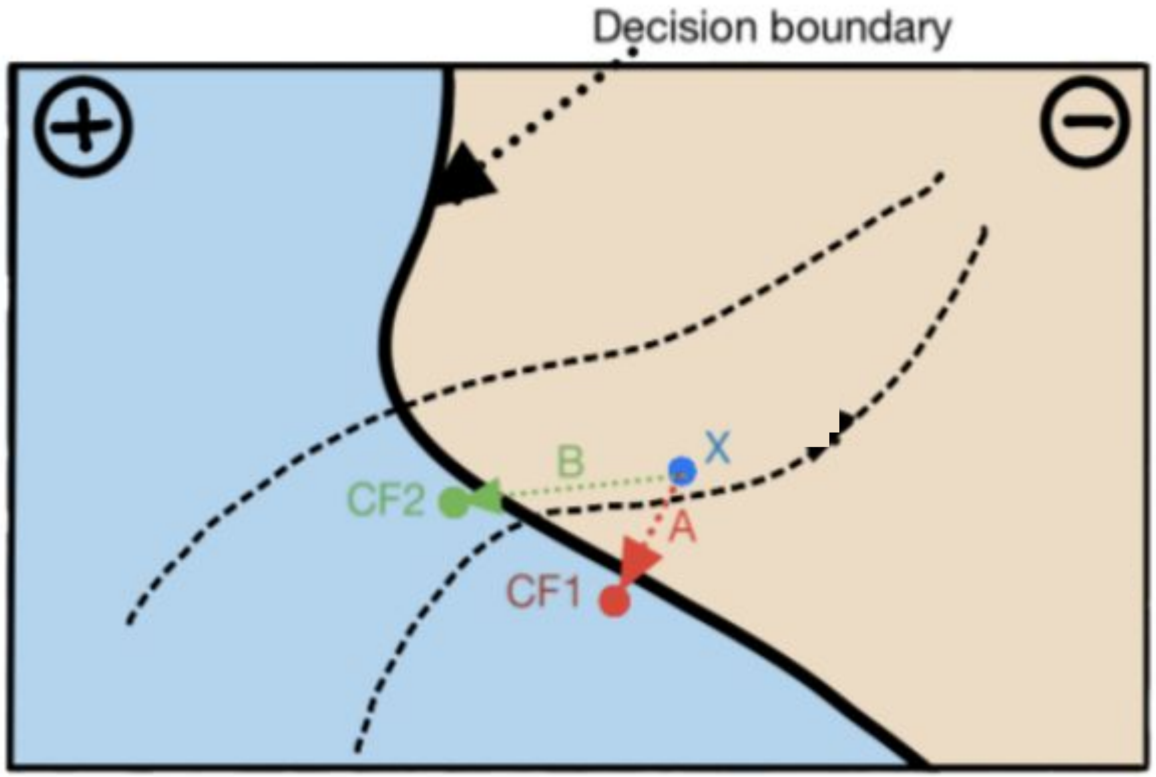
Counterfactual Explanations
What features need to be changed and by how much to flip a model's prediction ?
(i.e., to reverse an unfavorable outcome).

Counterfactual Explanations



Recourse: Increase your salary by 5K & pay your credit card bills on time for next 3 months

Generating Counterfactual Explanations: Intuition



Proposed solutions differ on:

1. How to choose among candidate counterfactuals?

1. How much access is needed to the underlying predictive model?

Take 1: Minimum Distance Counterfactuals

Distance Metric

Counterfactual

$$\arg\min_{x'} d(x, x')$$

Original Instance

$$s.t.\ f(x') = y'$$

Predictive Model

Desired Outcome

Choice of distance metric dictates what kinds of counterfactuals are chosen.

Wachter et. al. use normalized Manhattan distance.

[Wachter et. al., 2018]

Take 1: Minimum Distance Counterfactuals

$$\arg\min_{x'} d(x, x') \quad s.t. \; f(x') = y'$$

$$\arg\min_{x'} \lambda \, (f(x') - y')^2 + d(x, x')$$

Wachter et. al. solve a differentiable, unconstrained version of the objective using ADAM optimization algorithm with random restarts.

This method ***requires access to gradients*** of the underlying predictive model.

Take 1: Minimum Distance Counterfactuals

Person 1: If your LSAT was 34.0, you would have an average predicted score (0).

Person 2: If your LSAT was 32.4, you would have an average predicted score (0).

Person 3: If your LSAT was 33.5, and you were ‘white’, you would have an average predicted score (0).

Person 4: If your LSAT was 35.8, and you were ‘white’, you would have an average predicted score (0).

Person 5: If your LSAT was 34.9, you would have an average predicted score (0).

Not feasible to act upon these features!

[Ustun et. al., 2019]

Take 2: Feasible and Least Cost Counterfactuals

$$\arg\min_{x'} d(x, x') \quad s.t.\ f(x') = y'$$

$$\Rightarrow$$

$$\arg\min_{x' \in \boxed{\mathcal{A}}} \boxed{cost}(x, x') \quad s.t.\ f(x') = y'$$

- $\mathcal{A}$ is the set of feasible counterfactuals (input by end user)
  - E.g., changes to race, gender are not feasible
- *Cost* is modeled as total log-percentile shift
  - Changes become harder when starting off from a higher percentile value

Take 2: Feasible and Least Cost Counterfactuals

$$\arg\min_{x'} d(x, x') \quad s.t.\ f(x') = y'$$

$$\arg\min_{x' \in \mathcal{A}} cost(x, x') \quad s.t.\ f(x') = y'$$

- Ustun et. al. only consider the case where the model is a linear classifier
  - Objective formulated as an IP and optimized using CPLEX

- Requires ***complete access*** to the linear classifier i.e., weight vector

[Ustun et. al., 2019]

Take 2: Feasible and Least Cost Counterfactuals

$$\arg\min_{x'} d(x, x') \quad s.t. \; f(x') = y'$$

$$\arg\min_{x' \in \mathcal{A}} cost(x, x') \quad s.t. \; f(x') = y'$$

Question: What if we have a black box or a non-linear classifier?

Answer: generate a local linear model approximation (e.g., using LIME) and then apply Ustun et. al.'s framework

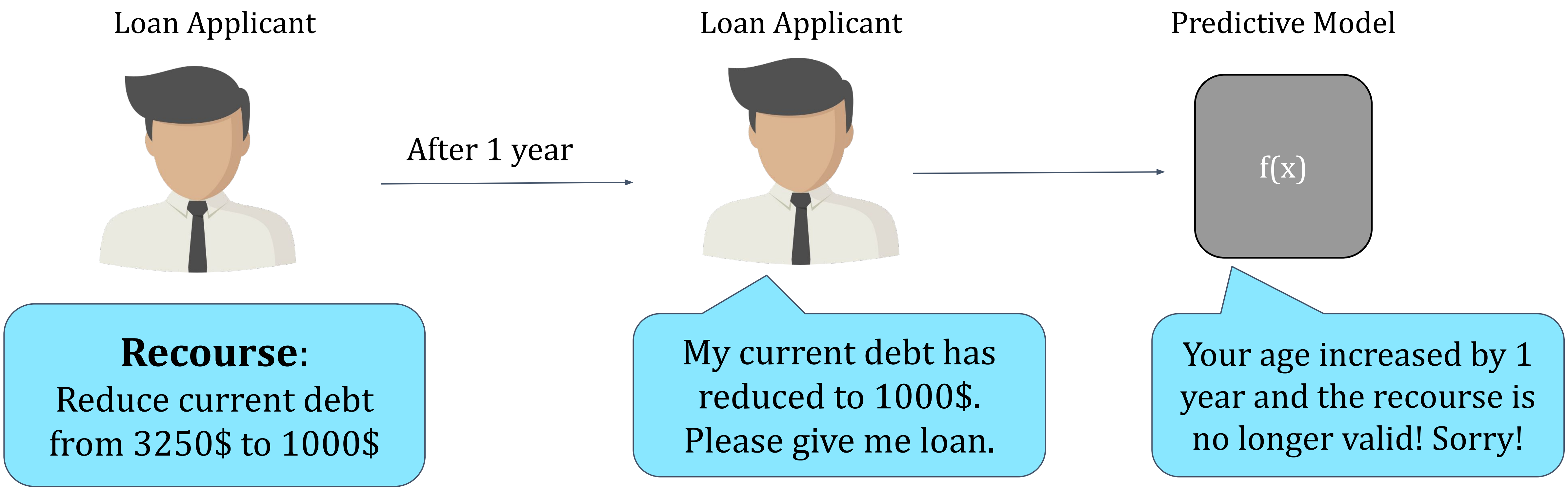
[Ustun et. al., 2019]

Take 2: Feasible and Least Cost Counterfactuals

Features to Change	Current Values		Required Values
n_credit_cards	5	⟶	3
current_debt	$3,250	⟶	$1,000
has_savings_account	FALSE	⟶	TRUE
has_retirement_account	FALSE	⟶	TRUE

Changing one feature without affecting another might not be possible!

Take 3: Causally Feasible Counterfactuals



Important to account for *feature interactions* when generating counterfactuals!

But how?!

Take 3: Causally Feasible Counterfactuals

$$\arg\min_{x' \in \mathcal{A}} cost(x, x')$$

$$s.t. \; f(x') = y'$$

$\mathcal{A}$ is the set of causally feasible counterfactuals permitted according to a given Structural Causal Model (SCM).

Question: What if we don't have access to the structural causal model?

[Verma et. al., 2020, Pawelczyk et. al., 2020]

Counterfactuals on Data Manifold

- Generated counterfactuals should lie on the data manifold
- Construct Variational Autoencoders (VAEs) to map input instances to latent space
- Search for counterfactuals in the latent space
- Once a counterfactual is found, map it back to the input space using the decoder

Approaches for Post hoc Explainability

Local Explanations

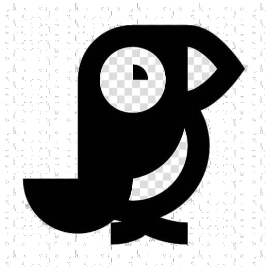
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Global Explanations

- Explain the complete behavior of a given (black box) model
  - Provide a *bird's eye view* of model behavior



- Help detect *big picture* model biases persistent across larger subgroups of the population
  - Impractical to manually inspect local explanations of several instances to ascertain big picture biases!

- Global explanations are complementary to local explanations

Local vs. Global Explanations

Explain individual predictions

Explain complete behavior of the model

Help unearth biases in the *local neighborhood* of a given instance

Help shed light on *big picture biases* affecting larger subgroups

Help vet if individual predictions are being made for the right reasons

Help vet if the model, at a high level, is suitable for deployment

Approaches for Post hoc Explainability

Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Global Explanation as a Collection of Local Explanations

How to generate a global explanation of a (black box) model?

- Generate a local explanation for every instance in the data using one of the approaches discussed earlier

- Pick a subset of *k* local explanations to constitute the global explanation

What local explanation technique to use?
How to choose the subset of k local explanations?

Global Explanations from Local Feature Importances: SP-LIME

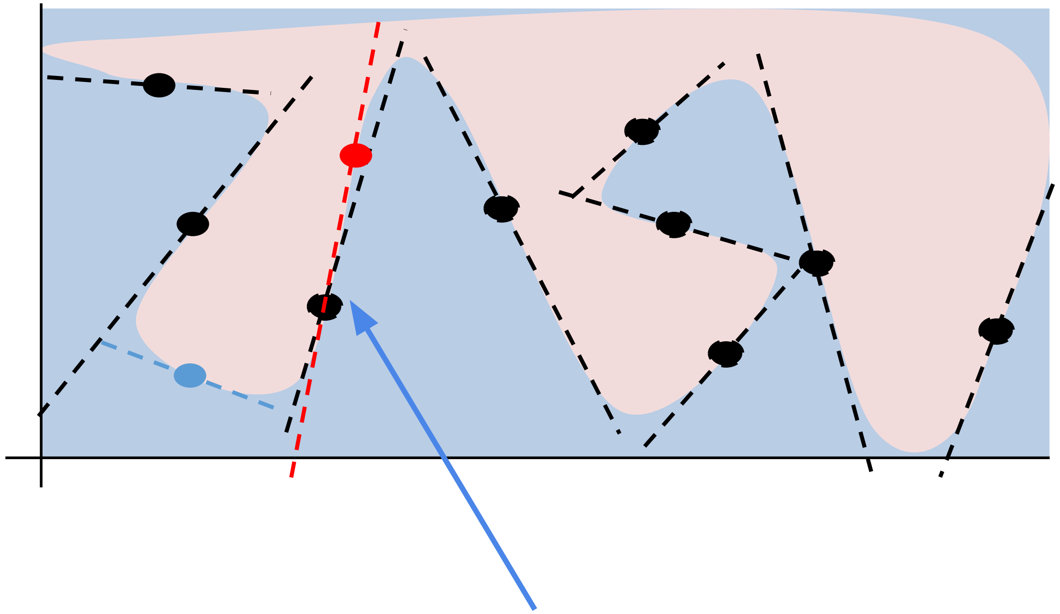
LIME explains a single prediction
local behavior for a single instance

Can't examine all explanations
Instead pick *k* explanations to show to the user

Representative
Should summarize the model's global behavior

Diverse
Should not be redundant in their descriptions

SP-LIME uses submodular optimization and *greedily* picks k explanations



Model Agnostic

Global Explanations from Local Rule Sets: SP-Anchor

- Use Anchors algorithm discussed earlier to obtain local rule sets for every instance in the data

- Use the same procedure to *greedily select a subset of k local rule sets* to correspond to the global explanation

Approaches for Post hoc Explainability

Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

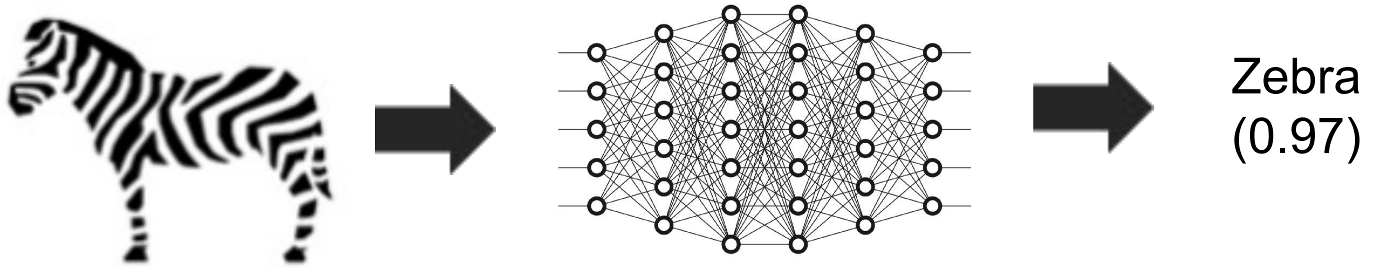
- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Representation Based Approaches

- Derive model understanding by analyzing intermediate representations of a DNN.

- Determine model's reliance on 'concepts' that are semantically meaningful to humans.

Representation Based Explanations



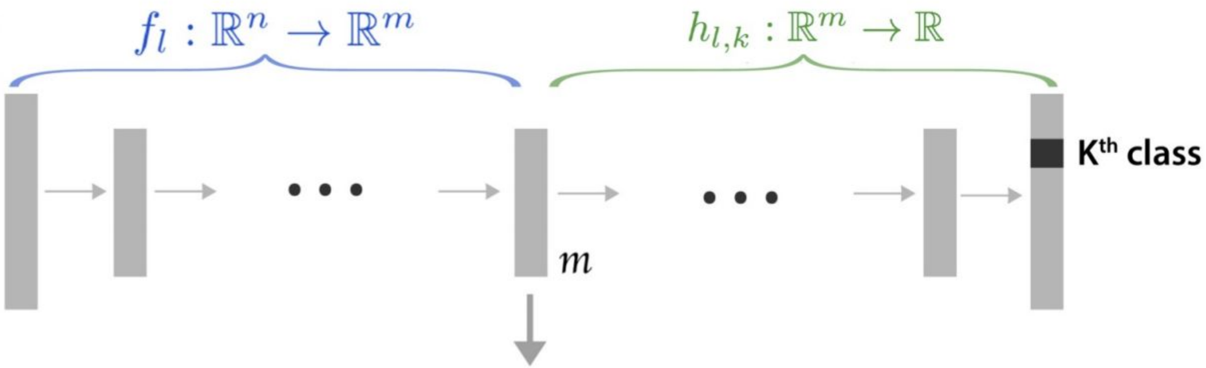
How important is the notion of “stripes” for this prediction?

[Kim et. al., 2018]

Representation Based Explanations: TCAV

Examples of the concept “stripes”

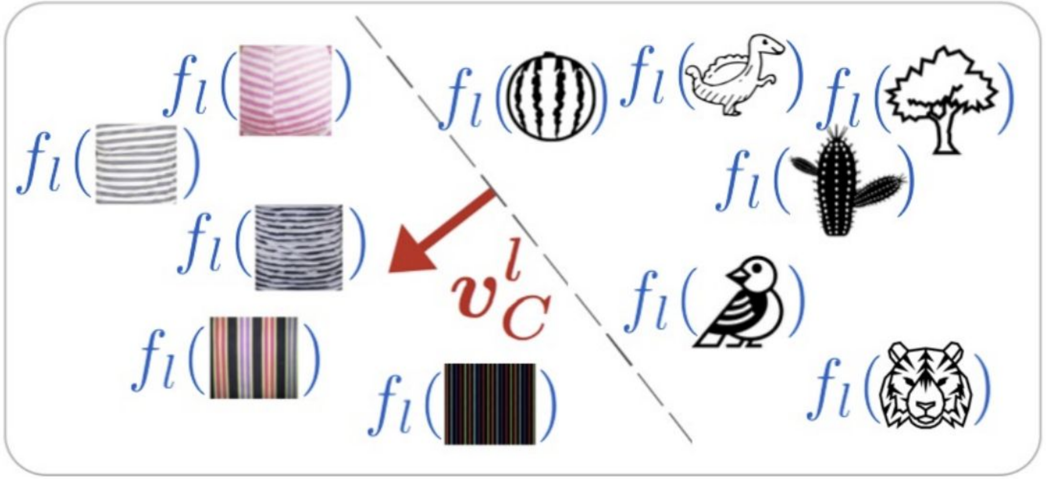
Random examples



Train a linear classifier to separate activations

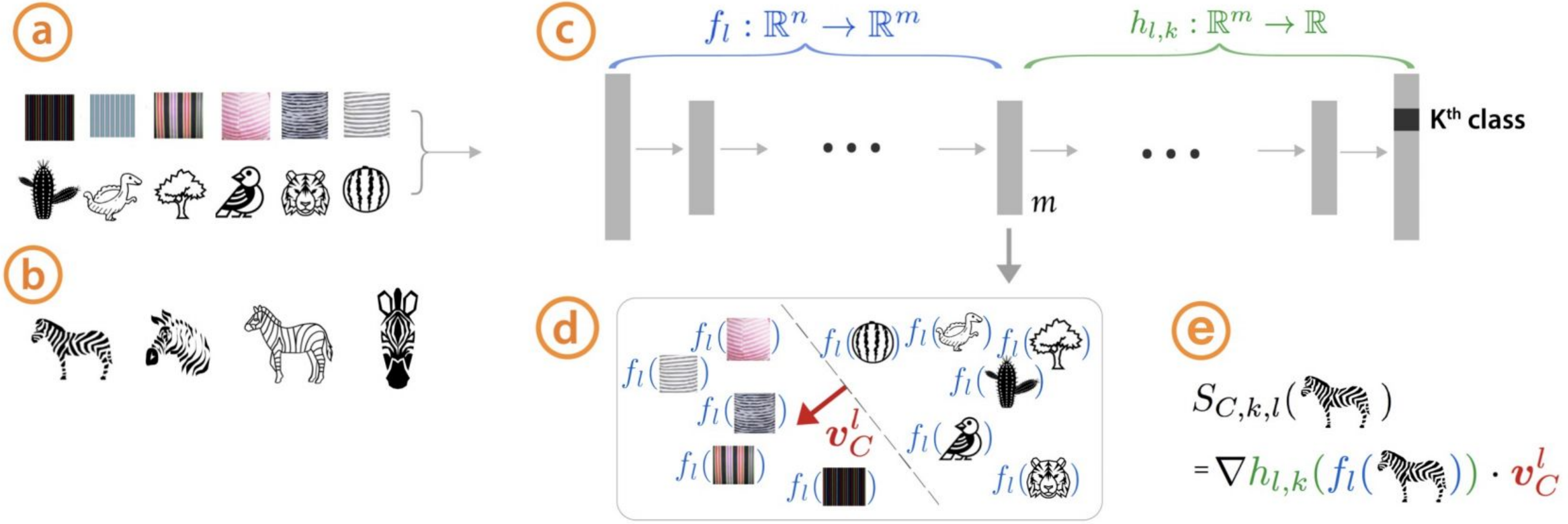
The vector orthogonal to the decision boundary pointing towards the “stripes” class quantifies the concept “stripes”

Compute derivatives by leveraging this vector to determine the importance of the notion of stripes for any given prediction



Quantitative Testing with Concept Activation Vectors (TCAV)

TCAV measures the sensitivity of a model's prediction to **user provided concept** using the model **internal representations**.



Kim et. al. 2018

Approaches for Post hoc Explainability

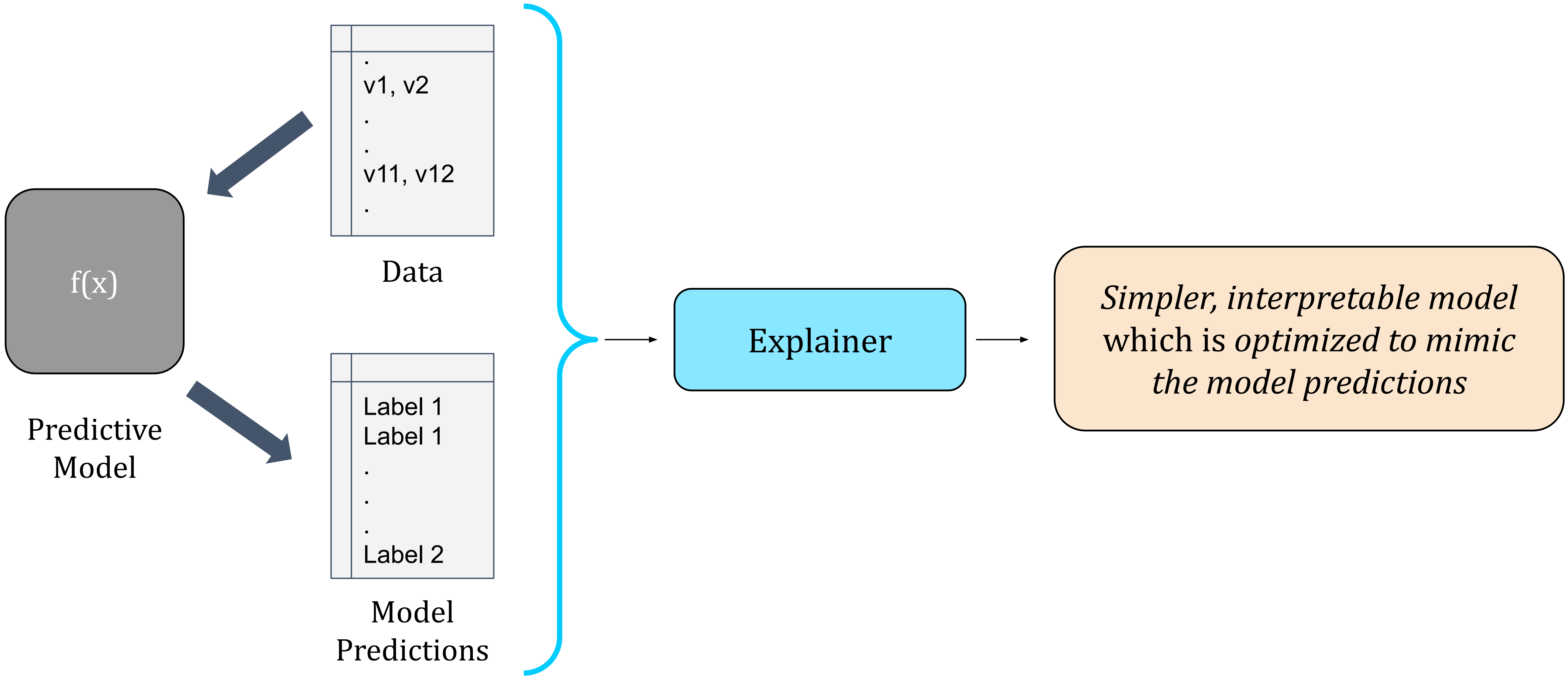
Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

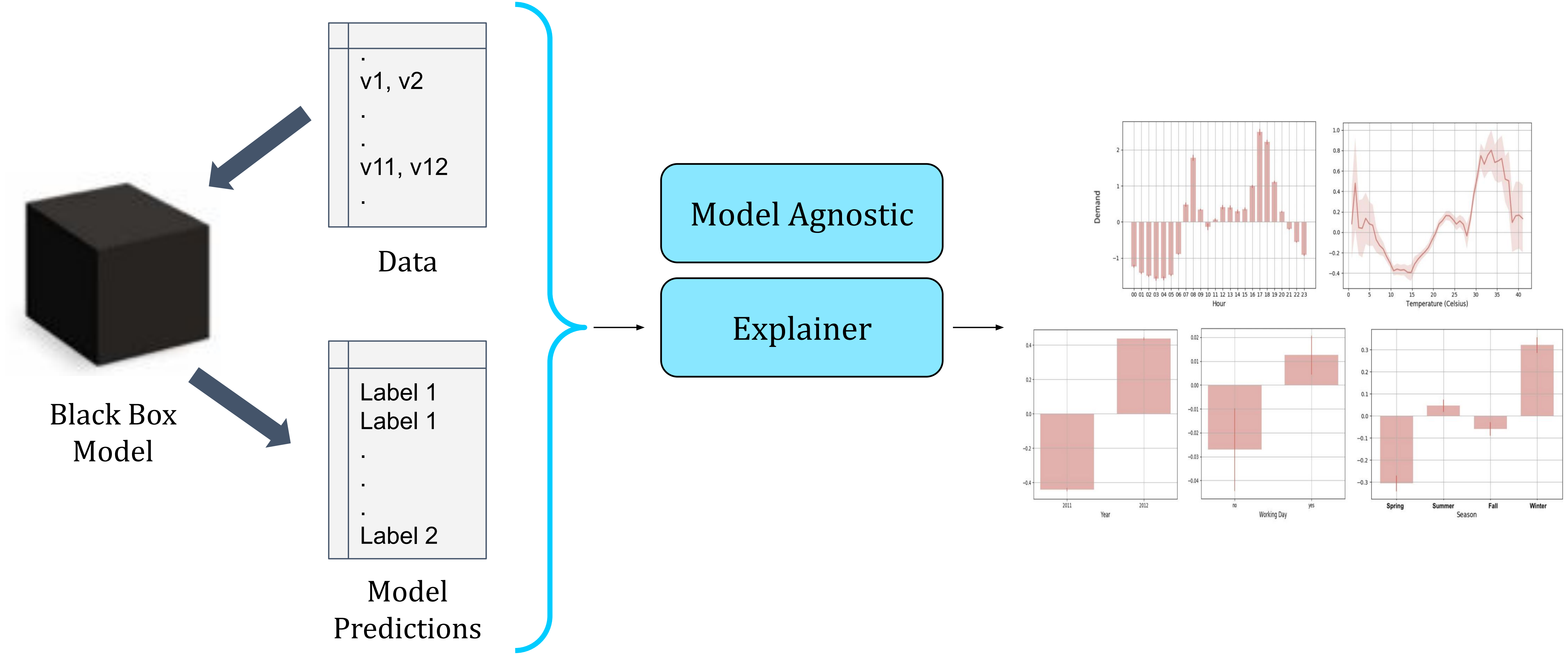
Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Model Distillation for Generating Global Explanations

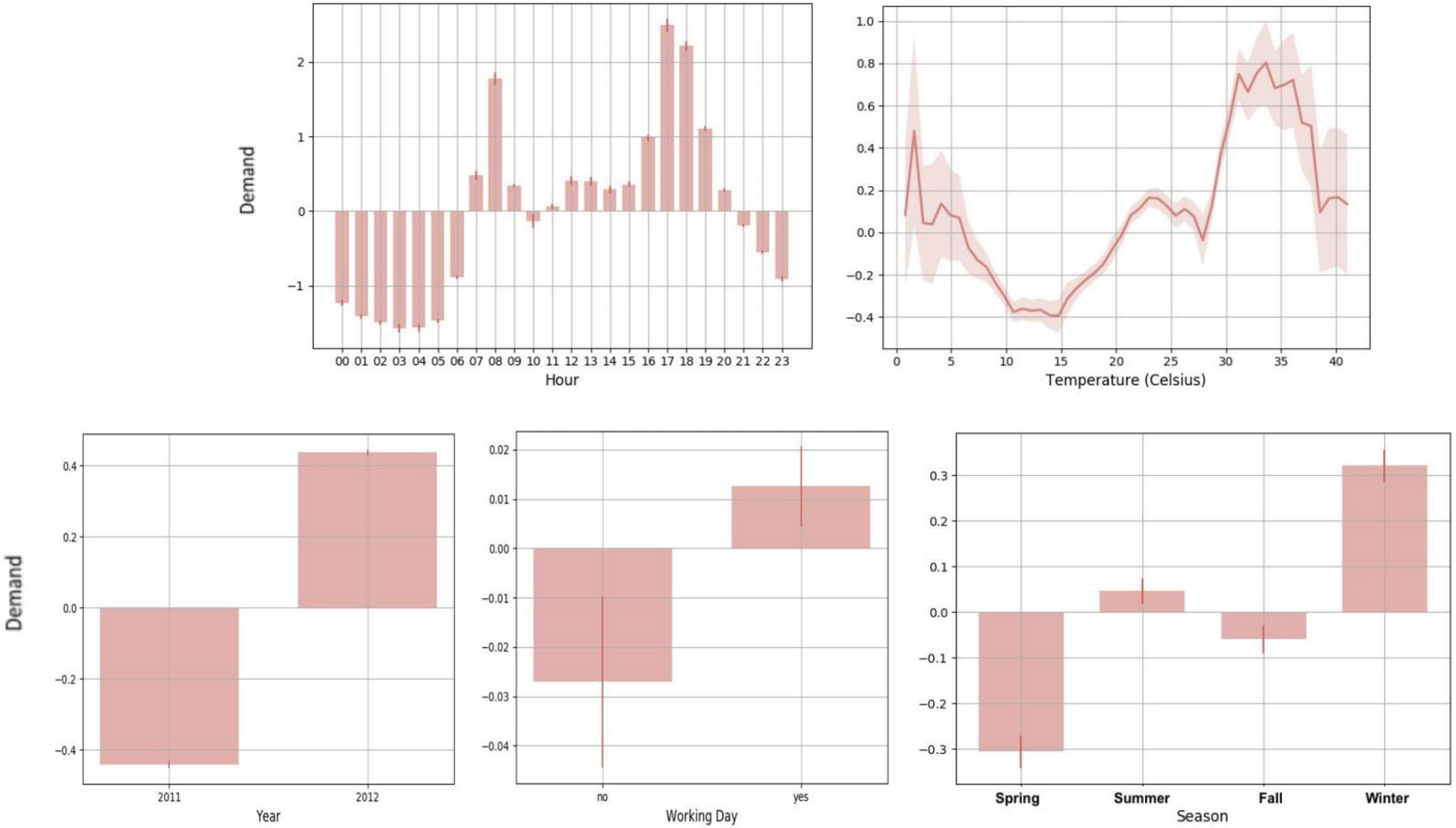


Generalized Additive Models as Global Explanations



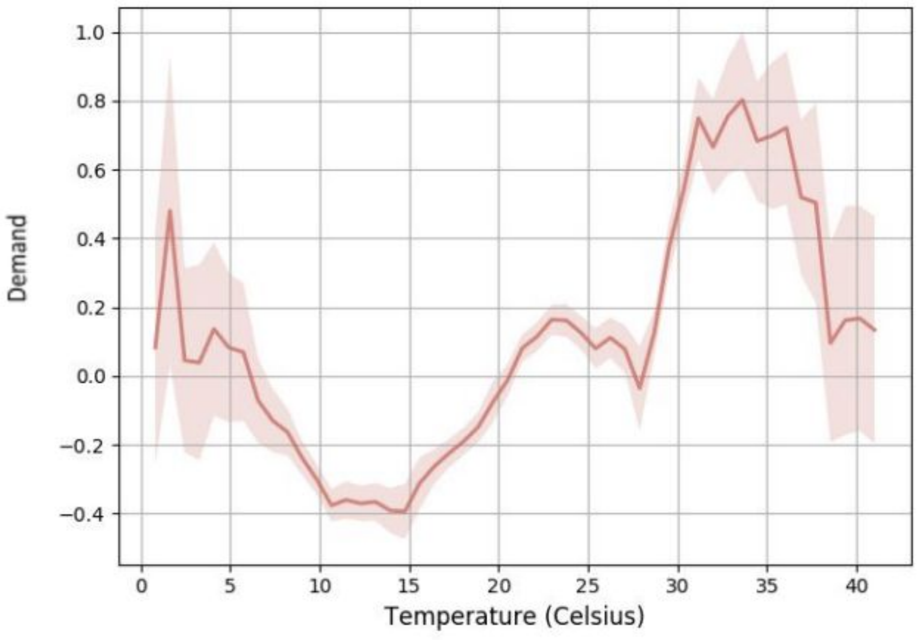
[Tan et. al., 2019]

Generalized Additive Models as Global Explanations: *Shape Functions* for Predicting Bike Demand



Generalized Additive Models as Global Explanations: *Shape Functions* for Predicting Bike Demand

How does bike demand vary as a function of temperature?



Generalized Additive Models as Global Explanations

Generalized Additive Model (GAM) :

$$\hat{y} = h_0 + \underbrace{\sum_i h_i(x_i)}_{\text{Shape functions of individual features}} + \underbrace{\sum_{i \neq j} h_{ij}(x_i, x_j) + \sum_{i \neq j} \sum_{j \neq k} h_{ijk}(x_i, x_j, x_k) + \cdots}_{\text{Higher order feature interaction terms}}$$

Fit this model to the predictions of the black box to obtain the shape functions.

[Bastani et. al., 2019]

Decision Trees as Global Explanations

.
v1, v2
.
.
v11, v12
.

Data

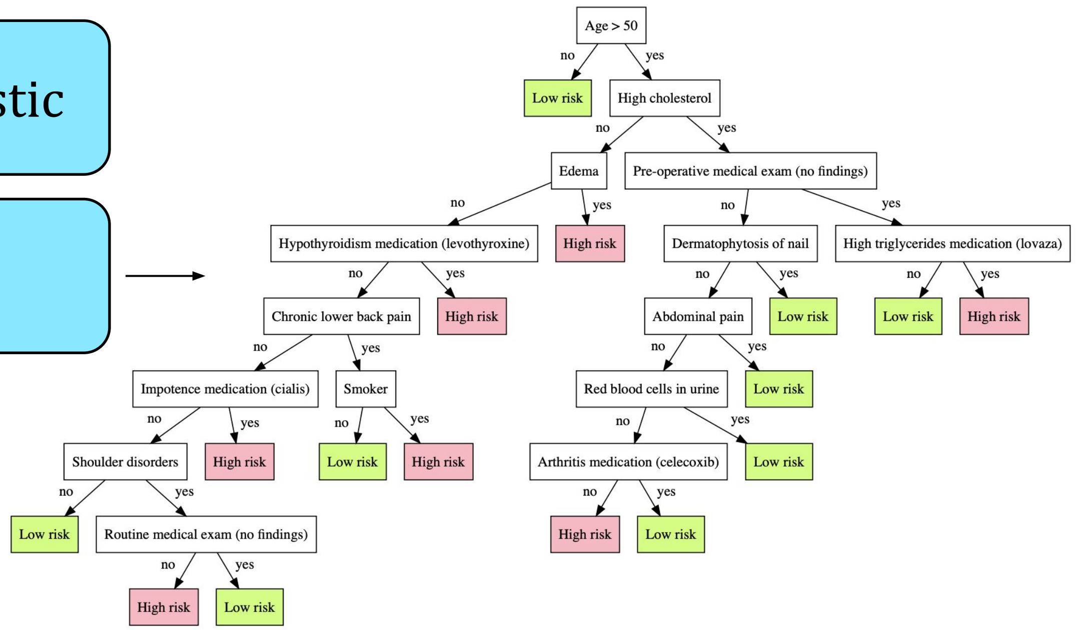
Black Box Model

Label 1
Label 1
.
.
.
Label 2

Model Predictions

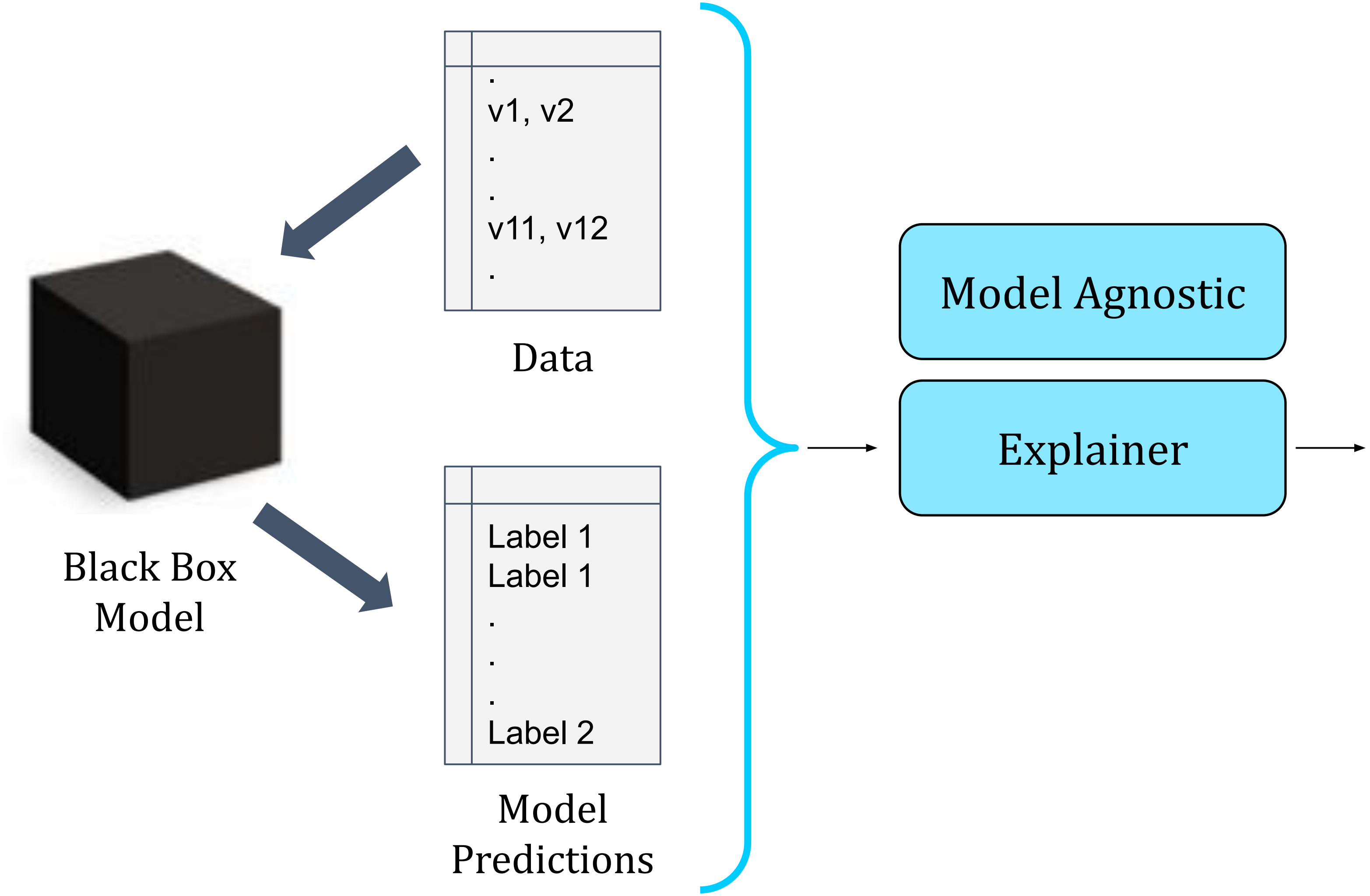
Model Agnostic

Explainer



[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations



If Age <50 **and** Male $=$Yes:

If Past-Depression $=$Yes **and** Insomnia $=$No **and** Melancholy $=$No, **then** Healthy

If Past-Depression $=$Yes **and** Insomnia $=$Yes **and** Melancholy $=$Yes **and** Tiredness $=$Yes, **then** Depression

If Age $\geq$ 50 **and** Male $=$No:

If Family-Depression $=$Yes **and** Insomnia $=$No **and** Melancholy $=$Yes **and** Tiredness $=$Yes, **then** Depression

If Family-Depression $=$No **and** Insomnia $=$No **and** Melancholy $=$No **and** Tiredness $=$No, **then** Healthy

Default:

If Past-Depression $=$Yes **and** Tiredness $=$No **and** Exercise $=$No **and** Insomnia $=$Yes, **then** Depression

If Past-Depression $=$No **and** Rapid-Weight-Gain $=$Yes **and** Tiredness $=$Yes **and** Melancholy $=$Yes, **then** Depression

[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations

Subgroup Descriptor

If Age <50 **and** Male $=$Yes:

If Past-Depression $=$Yes **and** Insomnia $=$No **and** Melancholy $=$No, **then** Healthy

If Past-Depression $=$Yes **and** Insomnia $=$Yes **and** Melancholy $=$Yes **and** Tiredness $=$Yes, **then** Depression

Decision Logic

If Age ≥ 50 **and** Male $=$No:

If Family-Depression $=$Yes **and** Insomnia $=$No **and** Melancholy $=$Yes **and** Tiredness $=$Yes, **then** Depression

If Family-Depression $=$No **and** Insomnia $=$No **and** Melancholy $=$No **and** Tiredness $=$No, **then** Healthy

Default:

If Past-Depression $=$Yes **and** Tiredness $=$No **and** Exercise $=$No **and** Insomnia $=$Yes, **then** Depression

If Past-Depression $=$No **and** Rapid-Weight-Gain $=$Yes **and** Tiredness $=$Yes **and** Melancholy $=$Yes, **then** Depression

[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations

If Exercise $=$Yes **and** Smoking $=$No:

If Rapid-Weight-Gain $=$Yes **and** Tiredness $=$Yes **and** Melancholy $=$Yes **and** Insomnia $=$Yes **and** Age $<$50, **then** Depression

If Tiredness $=$Yes **and** Melancholy $=$Yes **and** Age $\geq$ 50, **then** Depression

If Tiredness $=$No **and** Melancholy $=$No, **then** Healthy

If Smoking $=$Yes:

If Rapid-Weight-Gain $=$Yes **and** Melancholy $=$Yes, **then** Depression

If Tiredness $=$No **and** Insomnia $=$No **and** Melancholy $=$No **and** Rapid-Weight-Gain $=$No, **then** Healthy

If Insomnia $=$Yes **and** Past-Depression $=$Yes **and** Tiredness $=$Yes, **then** Depression

Default:

If Past-Depression $=$Yes **and** Tiredness $=$Yes **and** Melancholy $=$Yes, **then** Depression

If Past-Depression $=$No **and** Rapid-Weight-Gain $=$Yes **and** Tiredness $=$No **and** Melancholy $=$Yes, **then** Depression

If Family-Depression $=$Yes **and** Age $\geq$ 50 **and** Male $=$No **and** Tiredness $=$Yes, **then** Depression

[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations: Desiderata & Optimization Problem

Desiderata	Optimization
Fidelity Describe model behavior accurately	**Fidelity** Minimize number of instances for which explanation's label $\neq$ model prediction
Unambiguity No contradicting explanations	**Unambiguity** Minimize the number of duplicate rules applicable to each instance
Simplicity Users should be able to look at the explanation and reason about model behavior	**Simplicity** Minimize the number of conditions in rules; Constraints on number of rules & subgroups;
Customizability Users should be able to understand model behavior across various subgroups of interest	**Customizability** Outer rules should only comprise of features of user interest (candidate set restricted)

[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations

- The complete optimization problem is *non-negative*, *non-normal*, *non-monotone*, and *submodular* with *matroid constraints*

- Solved using the well-known *smooth local search* algorithm (Feige et. al., 2007) with best known optimality guarantees.

Approaches for Post hoc Explainability

Local Explanations

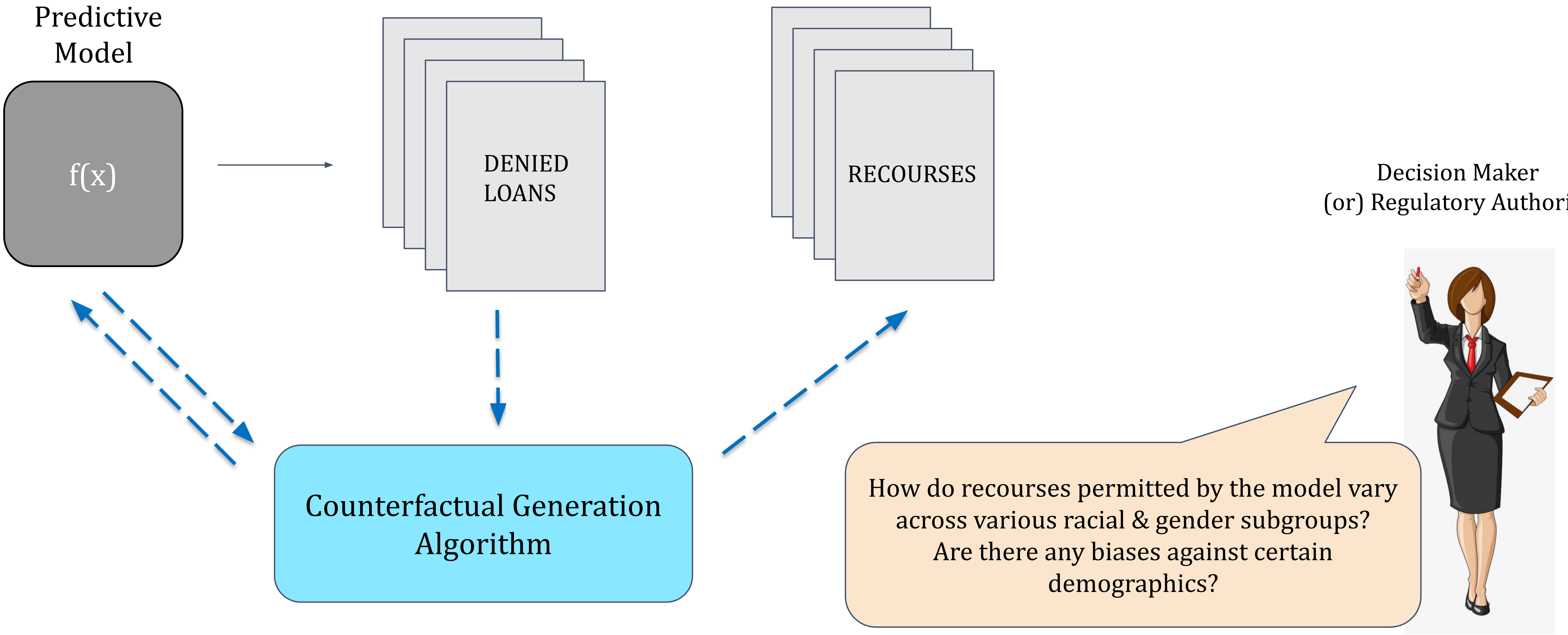
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

[Rawal and Lakkaraju, 2020]

Counterfactual Explanations



Customizable Global Summaries of Counterfactuals

Predictive Model

f(x)

DENIED LOANS

If Race = Caucasian **and** Gender = Male:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =No

If Race = Caucasian **and** Gender = Female:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =Yes

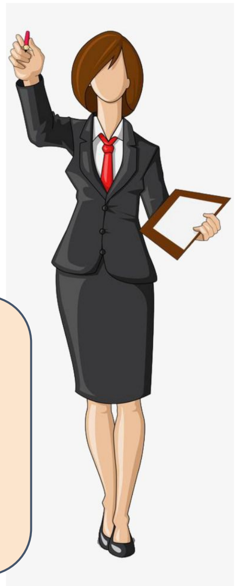
If Race ≠ Caucasian:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =Yes and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =Yes **and** Pays Rent =Yes

Algorithm for generating global summaries of counterfactuals

How do recourses permitted by the model vary across various racial & gender subgroups? Are there any biases against certain demographics?



[Rawal and Lakkaraju, 2020]

Customizable Global Summaries of Counterfactuals

Subgroup Descriptor

Omg! this model is biased. It requires certain demographics to “act upon” lot more features than others.

If Race = Caucasian **and** Gender = Male:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =No

If Race = Caucasian **and** Gender = Female:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =Yes

Recourse Rules

If Race $\neq$ Caucasian:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =Yes and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =Yes **and** Pays Rent =Yes

[Rawal and Lakkaraju, 2020]

Customizable Global Summaries of Counterfactuals: Desiderata & Optimization Problem

Desiderata	Optimization Problem
Recourse Correctness Prescribed recourses should obtain desirable outcomes	**Recourse Correctness** Minimize number of applicants for whom prescribed recourse does not lead to desired outcome
Recourse Coverage (Almost all) applicants should be provided with recourses	**Recourse Coverage** Minimize number of applicants for whom recourse does not exist (i.e., satisfy no rule).
Minimal Recourse Costs Acting upon a prescribed recourse should not be impractical or terribly expensive	**Minimal Recourse Costs** Minimize total ***feature costs*** as well as ***magnitude of changes*** in feature values
Interpretability of Summaries Summaries should be readily understandable to stakeholders (e.g., decision makers/regulatory authorities).	**Interpretability of Summaries** Constraints on # of rules, # of conditions in rules & # of subgroups
Customizability Stakeholders should be able to understand model behavior across various subgroups of interest	**Customizability** Outer rules should only comprise of features of stakeholder interest (candidate set restricted)

[Rawal and Lakkaraju, 2020]

Customizable Global Summaries of Counterfactuals

- The complete optimization problem is *non-negative*, *non-normal*, *non-monotone*, and *submodular* with *matroid constraints*

- Solved using the well-known *smooth local search* algorithm (Feige et. al., 2007) with best known optimality guarantees.

Breakout Groups

- What concepts/ideas/approaches from our morning discussion stood out to you ?

- We discussed different basic units of interpretation -- prototypes, rules, risk scores, shape functions (GAMs), feature importances
  - Are some of these more suited to certain data modalities (e.g., tabular, images, text) than others?

- What could be some potential vulnerabilities/drawbacks of inherently interpretable models and post hoc explanation methods?

- Given the diversity of the methods we discussed, how do we go about evaluating inherently interpretable models and post hoc explanation methods?

Agenda

- ❑ Inherently Interpretable Models
- ❑ Post hoc Explanation Methods
- ❑ Evaluating Model Interpretations/Explanations
- ❑ Empirically & Theoretically Analyzing Interpretations/Explanations
- ❑ Future of Model Understanding

Evaluating Model Interpretations/Explanations

- Evaluating the meaningfulness or correctness of explanations
  - Diverse ways of doing this depending on the type of model interpretation/explanation
- Evaluating the interpretability of explanations

[Doshi-Velez and Kim, 2017]

Evaluating Interpretability

Evaluating Interpretability

- Functionally-grounded evaluation: Quantitative metrics – e.g., number of rules, prototypes --> lower is better!

- Human-grounded evaluation: binary forced choice, forward simulation/prediction, counterfactual simulation

- Application-grounded evaluation: Domain expert with exact application task or simpler/partial task

Evaluating Inherently Interpretable Models

- Evaluating the accuracy of the resulting model
- Evaluating the interpretability of the resulting model
- Do we need to evaluate the “correctness” or “meaningfulness” of the resulting interpretations?

[Letham et. al. 2016]

Evaluating Bayesian Rule Lists

- A rule list classifier for stroke prediction

if hemiplegia **and** age >60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age >60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age ≤ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

[Lakkaraju et. al. 2016]

Evaluating Interpretable Decision Sets

- A decision set classifier for disease diagnosis

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure$\geq$ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI$\geq$ 0.3 **and** Insurance=None **and** Blood-Pressure$\geq$ 0.2 **then** Depression

If Smoker=Yes **and** BMI$\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

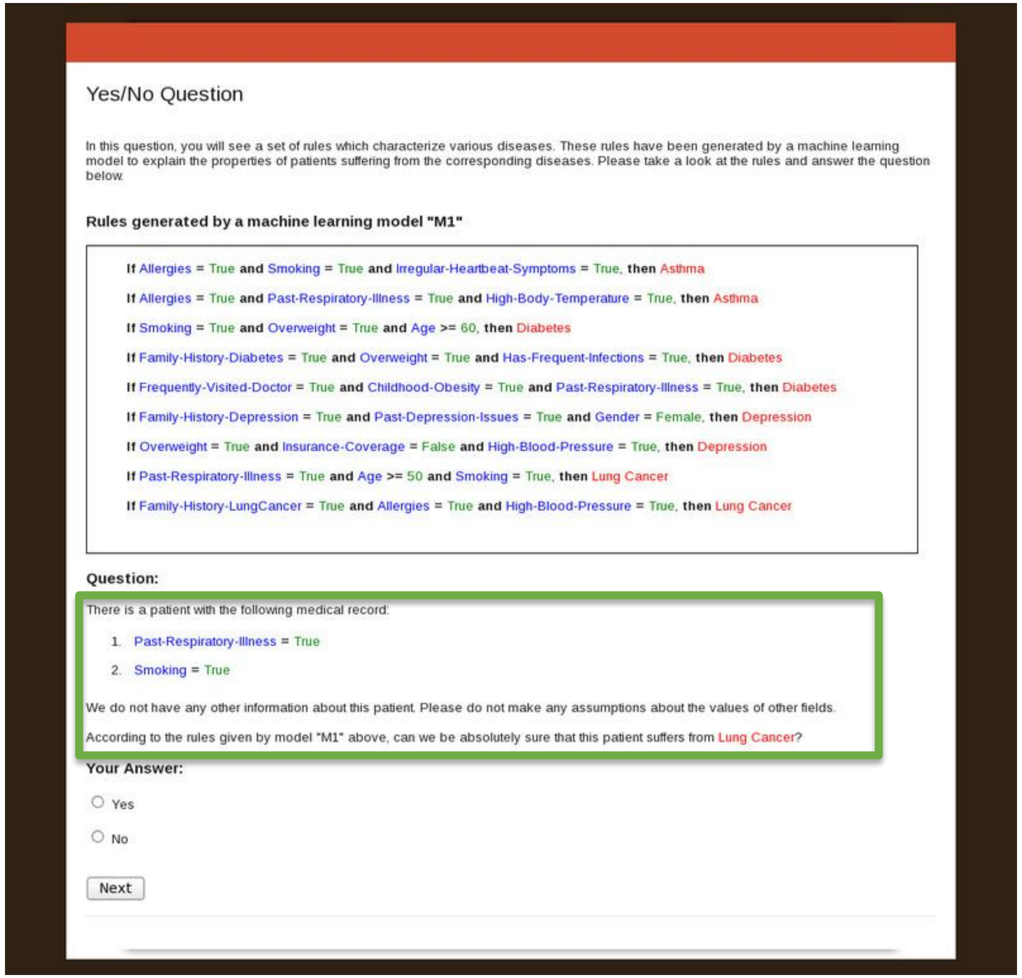
If Risk-Diabetes=Yes **and** BMI$\geq$ 0.4 **and** Prob-Infections$\geq$ 0.2 **then** Diabetes

If Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

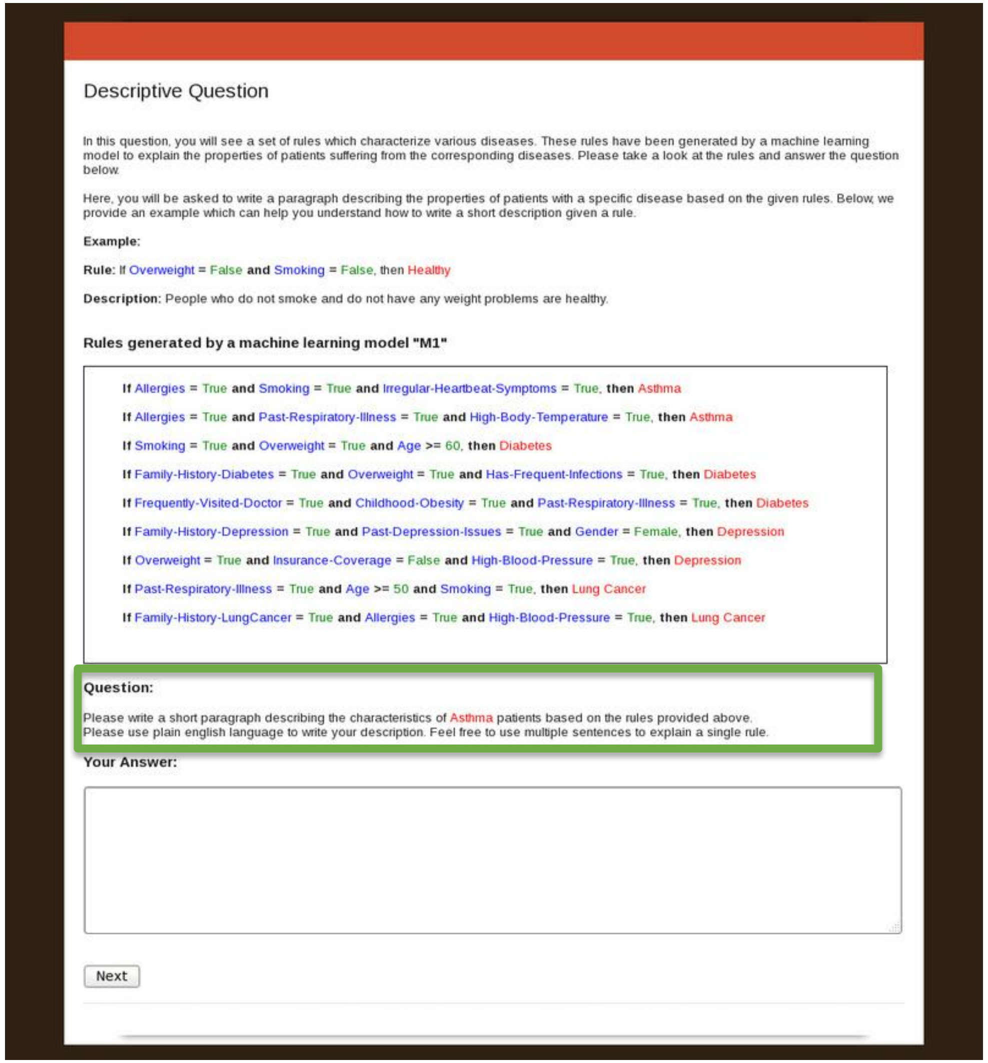
Evaluating Interpretability of Bayesian Rule Lists and Interpretable Decision Sets

- Number of rules, predicates etc. lower is better!
- User studies to compare interpretable decision sets to Bayesian Decision Lists (Letham et. al.)
- Each user is randomly assigned one of the two models
- 10 objective and 2 descriptive questions per user

Interface for Objective Questions



Interface for Descriptive Questions



Descriptive Question

In this question, you will see a set of rules which characterize various diseases. These rules have been generated by a machine learning model to explain the properties of patients suffering from the corresponding diseases. Please take a look at the rules and answer the question below.

Here, you will be asked to write a paragraph describing the properties of patients with a specific disease based on the given rules. Below, we provide an example which can help you understand how to write a short description given a rule.

Example:

Rule: If Overweight = False **and** Smoking = False, then Healthy

Description: People who do not smoke and do not have any weight problems are healthy.

Rules generated by a machine learning model "M1"

If Allergies = True **and** Smoking = True **and** Irregular-Heartbeat-Symptoms = True, **then** Asthma

If Allergies = True **and** Past-Respiratory-Illness = True **and** High-Body-Temperature = True, **then** Asthma

If Smoking = True **and** Overweight = True **and** Age >= 60, **then** Diabetes

If Family-History-Diabetes = True **and** Overweight = True **and** Has-Frequent-Infections = True, **then** Diabetes

If Frequently-Visited-Doctor = True **and** Childhood-Obesity = True **and** Past-Respiratory-Illness = True, **then** Diabetes

If Family-History-Depression = True **and** Past-Depression-Issues = True **and** Gender = Female, **then** Depression

If Overweight = True **and** Insurance-Coverage = False **and** High-Blood-Pressure = True, **then** Depression

If Past-Respiratory-Illness = True **and** Age >= 50 **and** Smoking = True, **then** Lung Cancer

If Family-History-LungCancer = True **and** Allergies = True **and** High-Blood-Pressure = True, **then** Lung Cancer

Question:

Please write a short paragraph describing the characteristics of Asthma patients based on the rules provided above.
Please use plain english language to write your description. Feel free to use multiple sentences to explain a single rule.

Your Answer:

Next

User Study Results

Task	Metrics	Our Approach	Bayesian Decision Lists
Descriptive	Human Accuracy	0.81	0.17
	Avg. Time Spent (secs.)	113.4	396.86
	Avg. # of Words	31.11	120.57
Objective	Human Accuracy	0.97	0.82
	Avg. Time Spent (secs.)	28.18	36.34

Objective Questions: 17% more accurate, 22% faster;
Descriptive Questions: 74% fewer words, 71% faster.

[Jain and Wallace, 2019]

Evaluating Prototype and Attention Layers

- Are prototypes and attention weights always meaningful?
- Do attention weights correlate with other measures of feature importance? E.g., gradients
- Would alternative attention weights yield different predictions?

No!!

[Agarwal et. al., 2022]

Evaluating Post hoc Explanations

- Evaluating the faithfulness (or correctness) of post hoc explanations

- Evaluating the stability of post hoc explanations

- Evaluating the fairness of post hoc explanations

- Evaluating the interpretability of post hoc explanations

Evaluating Faithfulness of Post hoc Explanations – Ground Truth

$$FeatureAgreement(E_a, E_b, k) = \frac{|top\_features(E_a, k) \cap top\_features(E_b, k)|}{k}$$

$$RankAgreement(E_a, E_b, k)$$

$$\frac{|\bigcup_{s \in S} \{s \mid s \in top\_features(E_a, k) \wedge s \in top\_features(E_b, k) \wedge rank(E_a, s) = rank(E_b, s)\}|}{k}$$

$$SignAgreement(E_a, E_b, k)$$

$$\frac{|\bigcup_{s \in S} \{s \mid s \in top\_features(E_a, k) \wedge s \in top\_features(E_b, k) \wedge sign(E_a, s) = sign(E_b, s)\}|}{k}$$

$$SignedRankAgreement(E_a, E_b, k)$$

$$\frac{\begin{array}{c}|\bigcup_{s \in S} \{s \mid s \in top\_features(E_a, k) \wedge s \in top\_features(E_b, k) \\ \wedge\, sign(E_a, s) = sign(E_b, s) \wedge rank(E_a, s) = rank(E_b, s)\}|\end{array}}{k}$$

Evaluating Faithfulness of Post hoc Explanations – Ground Truth

Spearman rank correlation coefficient computed over features of interest

$$RankCorrelation(E_a, E_b, F) = r_s(Ranking(E_a, F), Ranking(E_b, F))$$

$$PairwiseRankAgreement(E_a, E_b, F) = \frac{\sum_{i,j \text{ for } i<j} \mathbb{1}[RelativeRanking(E_a, f_i, f_j) = RelativeRanking(E_b, f_i, f_j)]}{\binom{|F|}{2}}$$

Evaluating Faithfulness of Post hoc Explanations – Explanations as Models

- If the explanation is itself a model (e.g., linear model fit by LIME), we can compute the fraction of instances for which the labels assigned by explanation model match those assigned by the underlying model

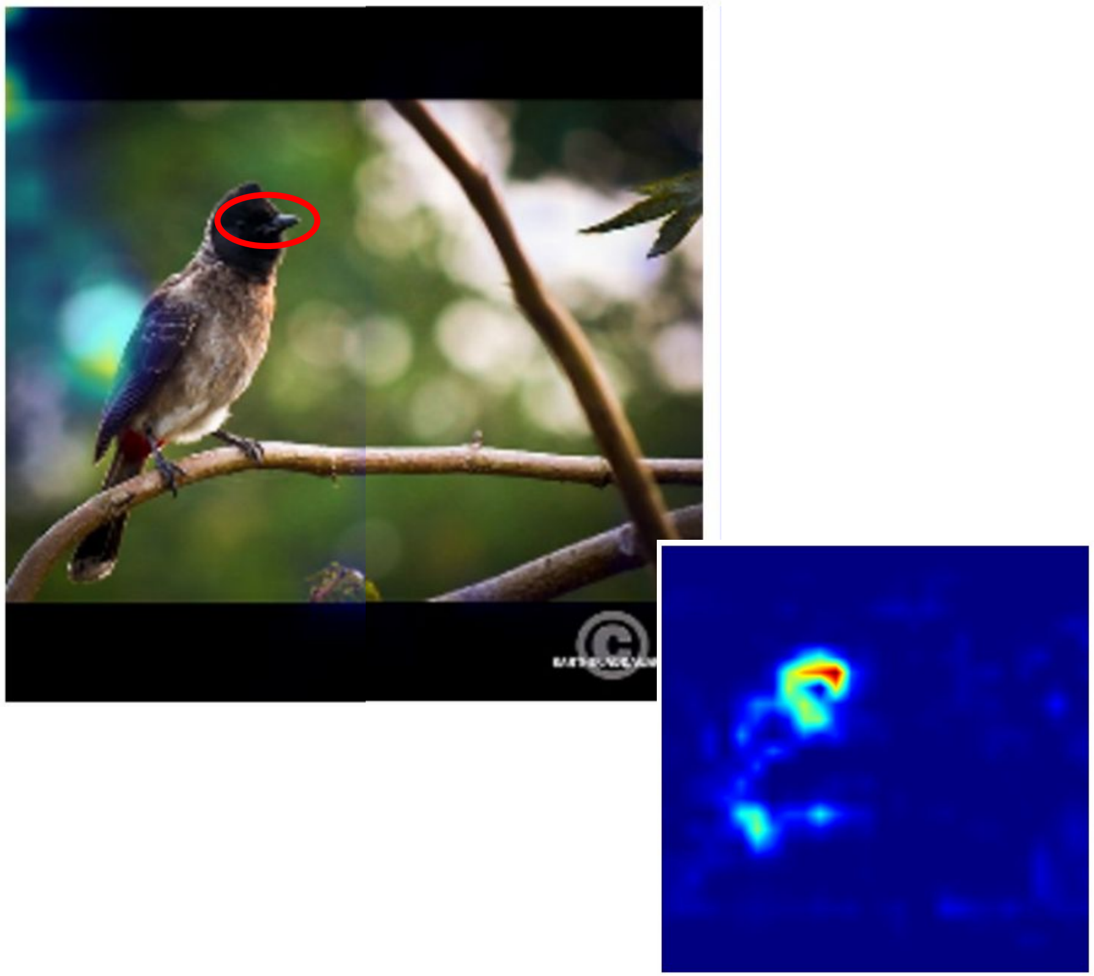
Evaluating Faithfulness of Post hoc Explanations

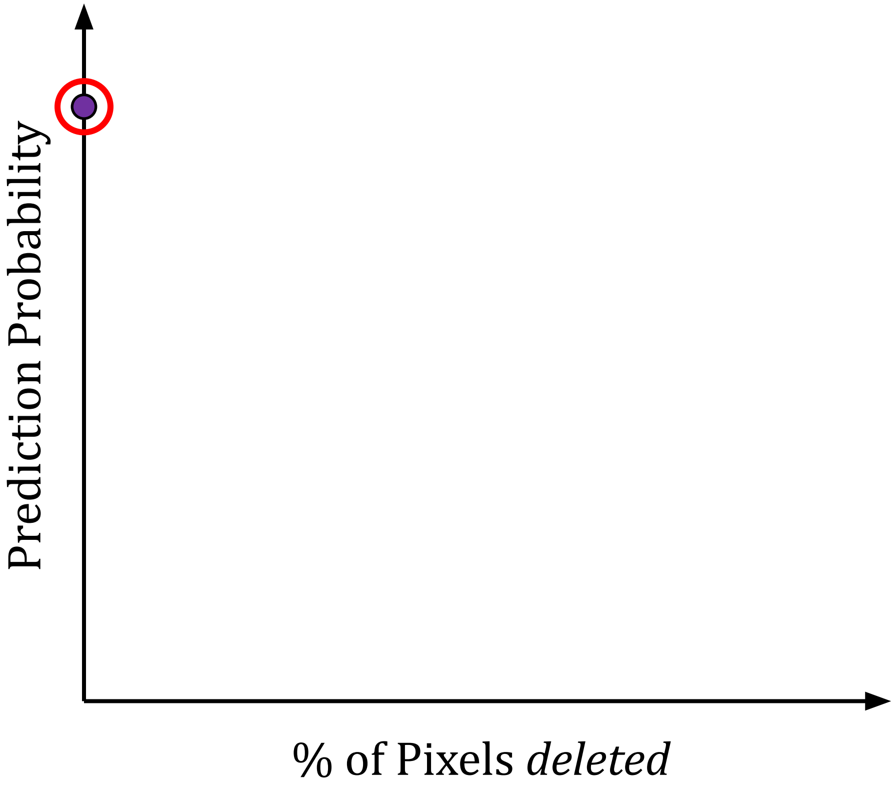
- What if we do not have any ground truth?

- What if explanations cannot be considered as models that output predictions?

How important are selected features?

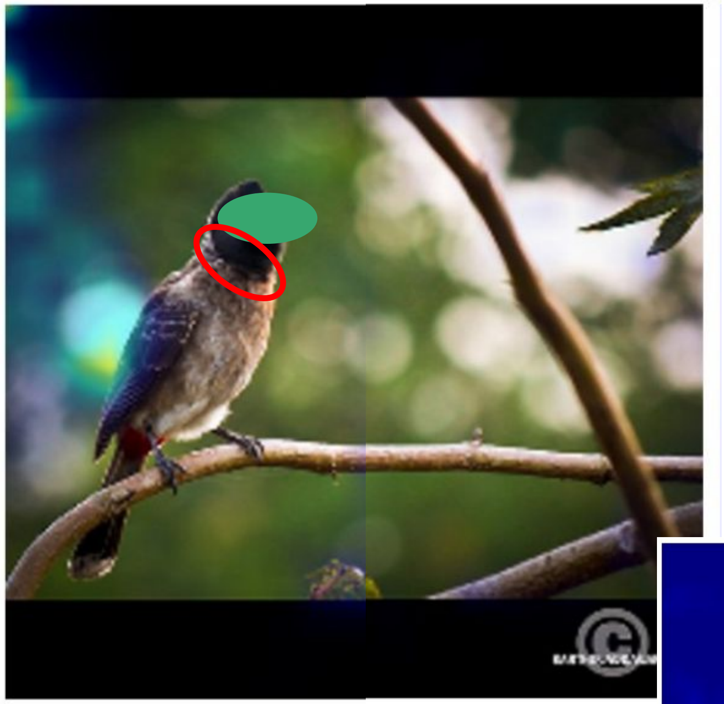
- Deletion: remove important features and see what happens..

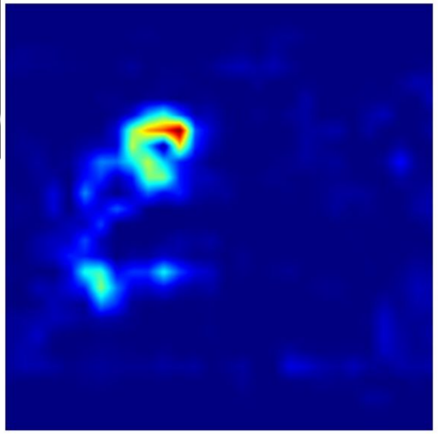


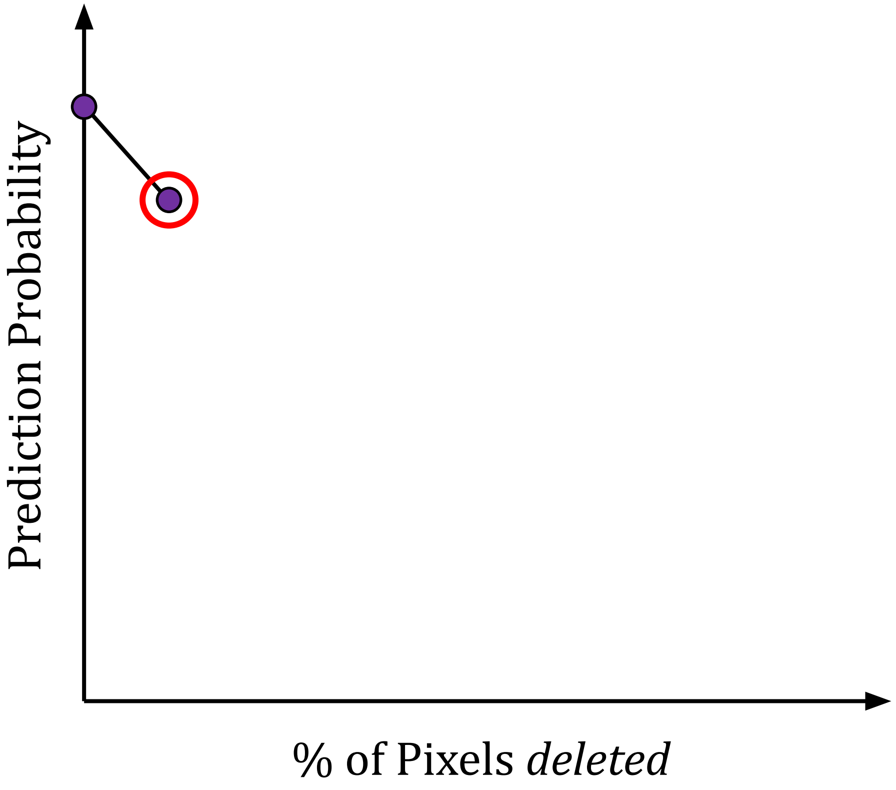


How important are selected features?

- Deletion: remove important features and see what happens..



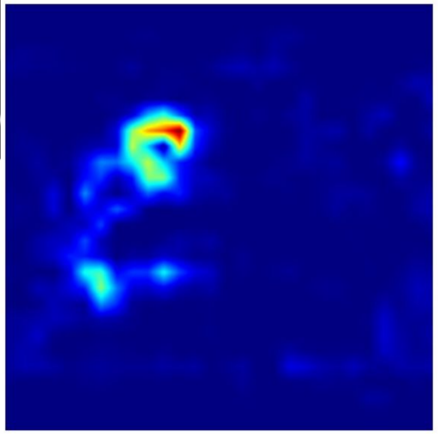


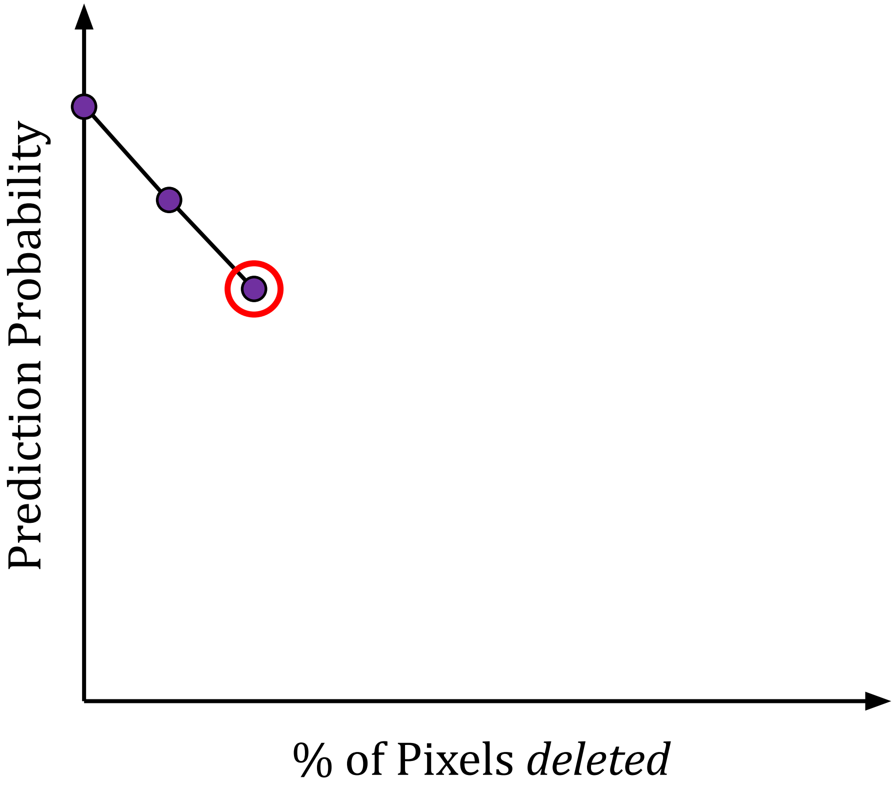


[Qi, Khorram, Fuxin, 2020]

How important are selected features?

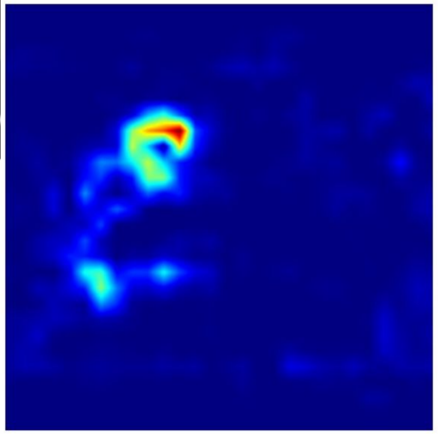
- Deletion: remove important features and see what happens..

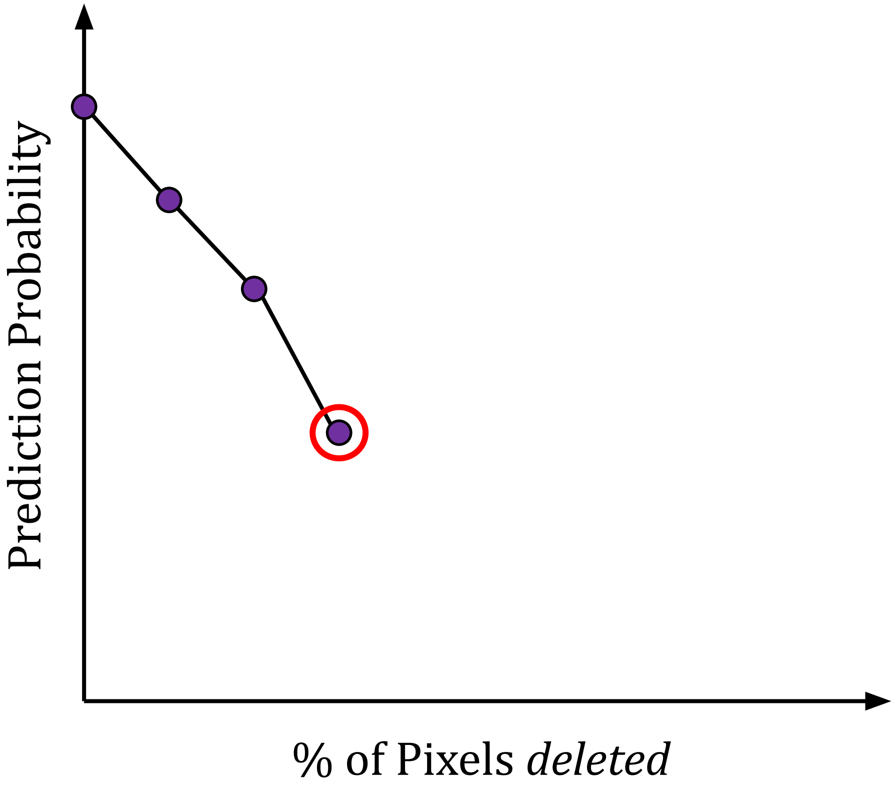




How important are selected features?

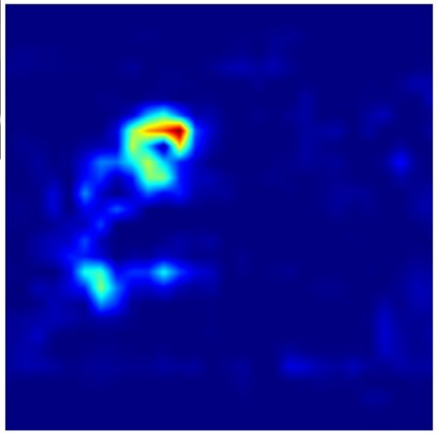
- Deletion: remove important features and see what happens..

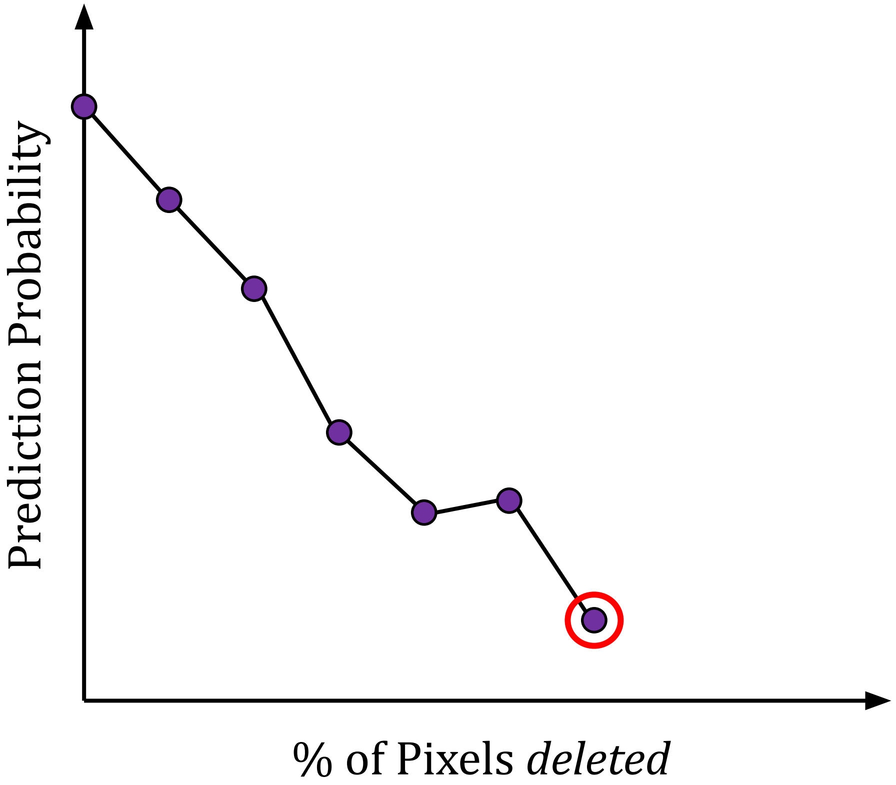




How important are selected features?

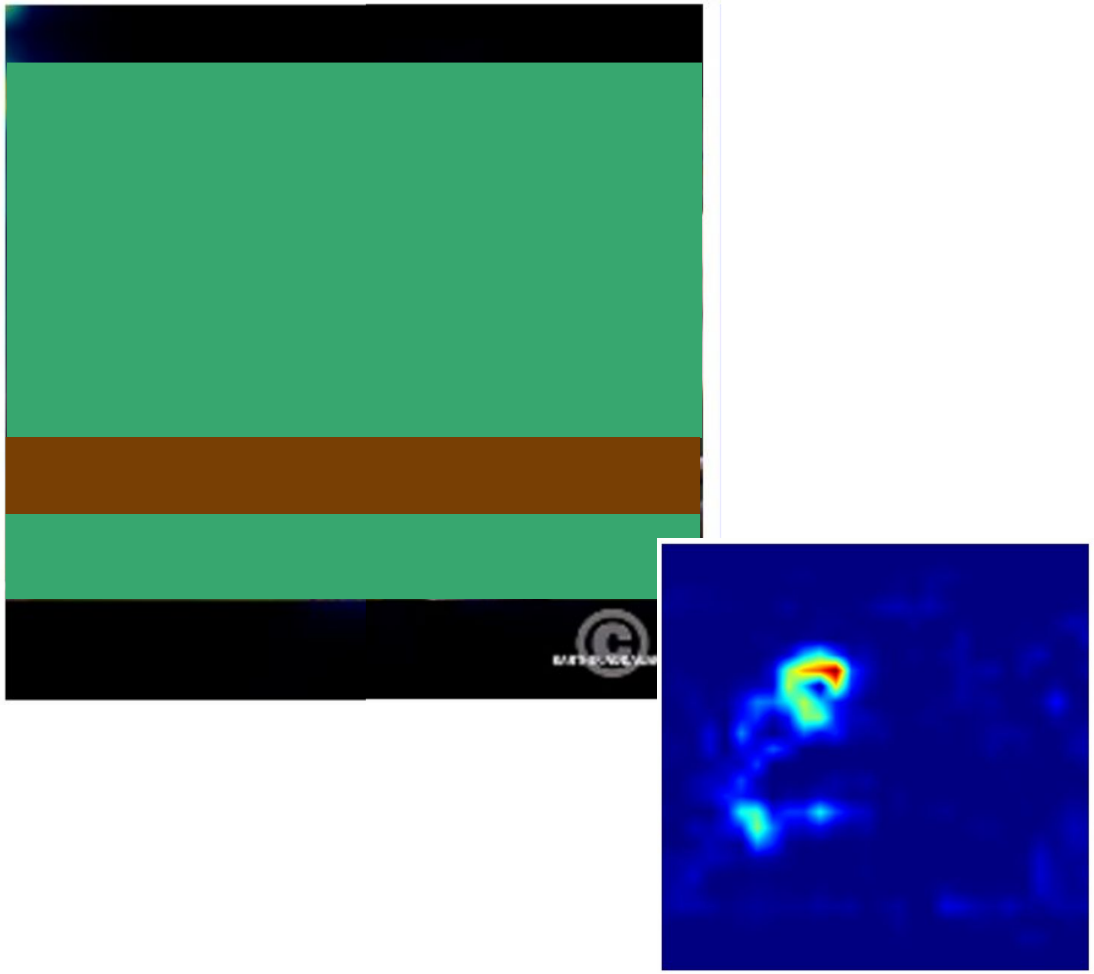
- Deletion: remove important features and see what happens..

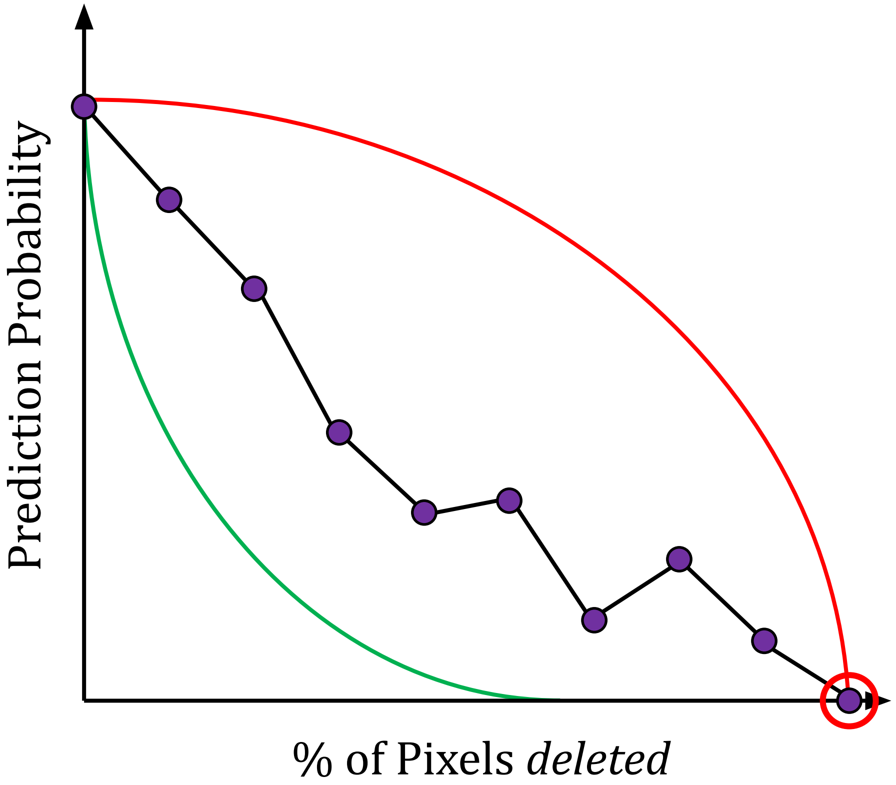




How important are selected features?

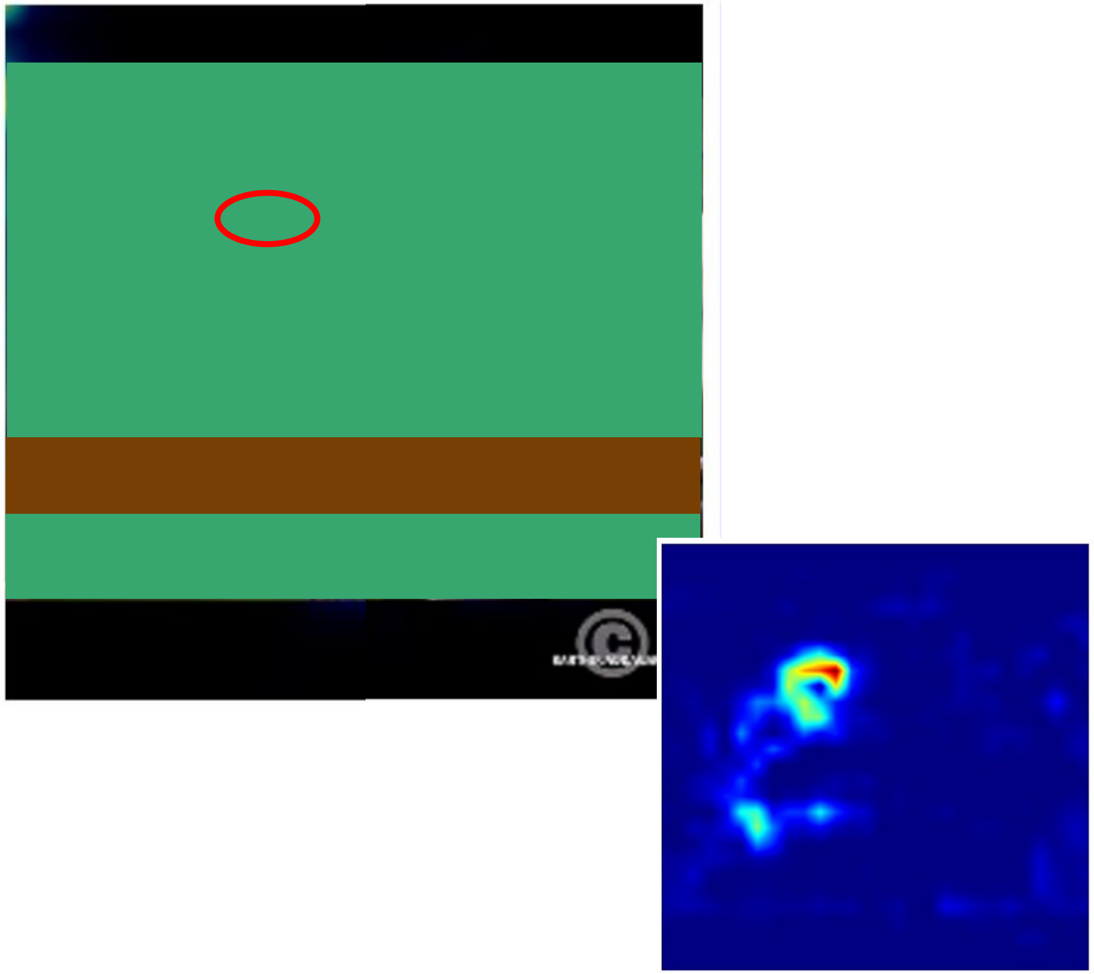
- Deletion: remove important features and see what happens..

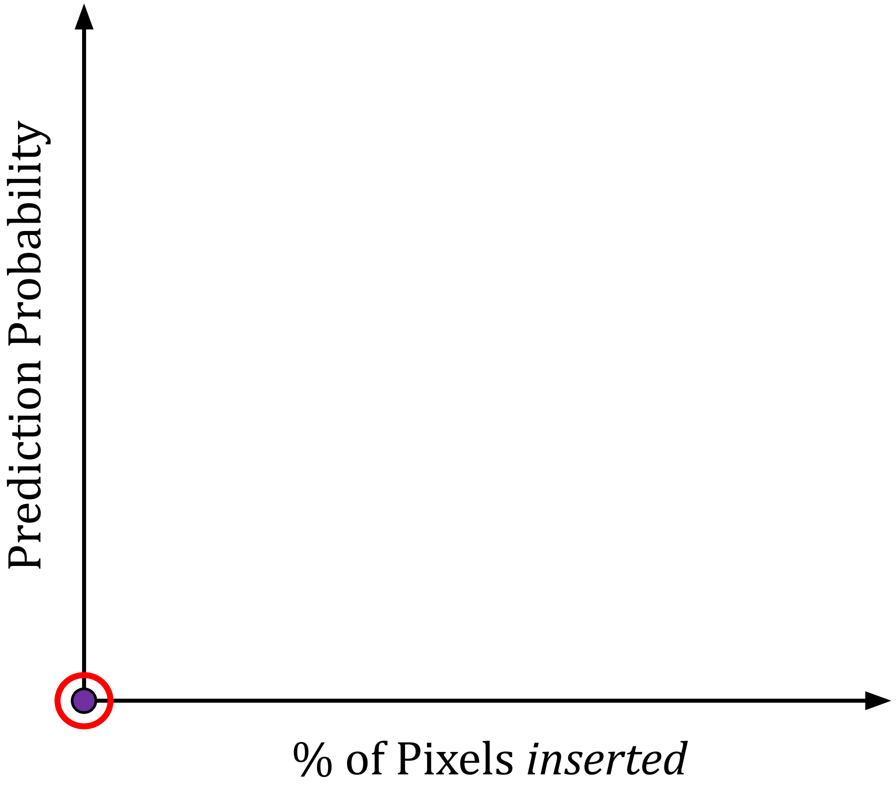




How important are selected features?

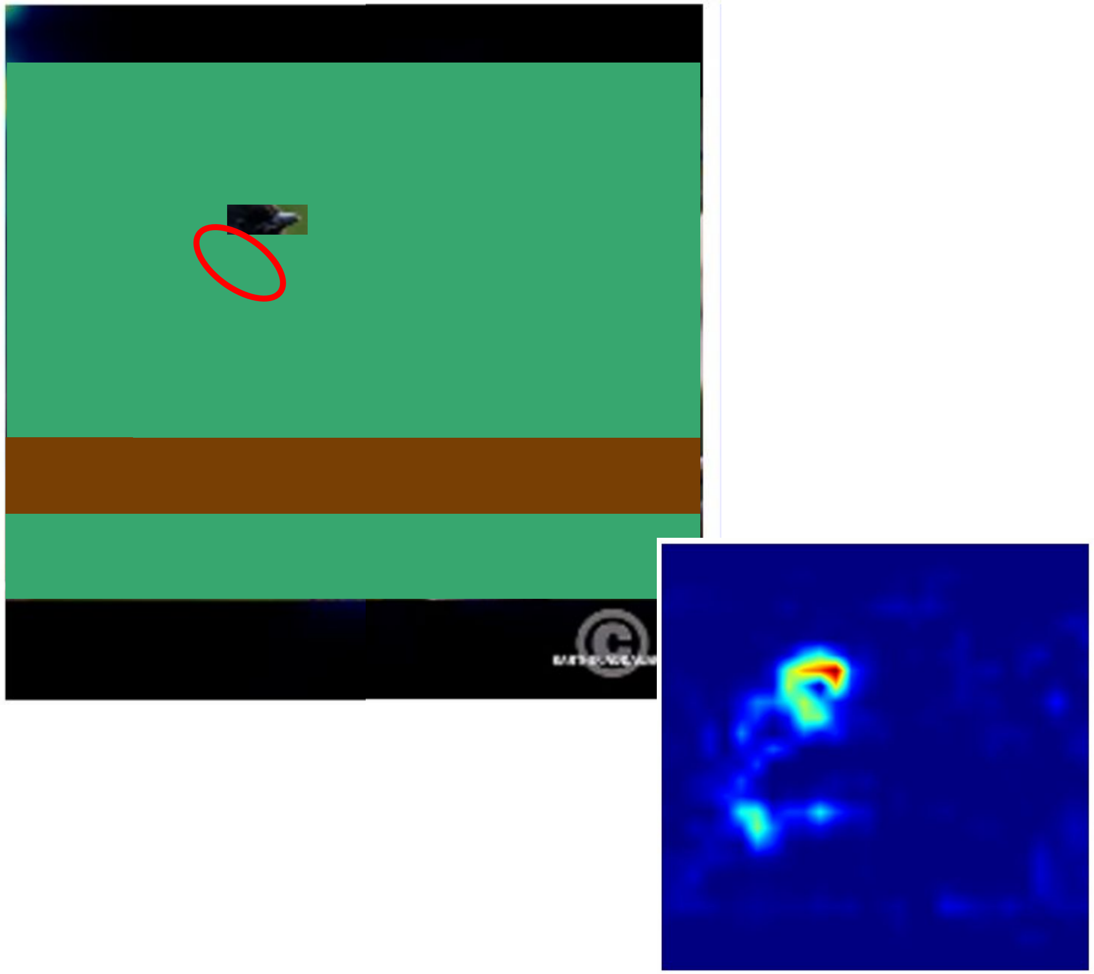
- Insertion: add important features and see what happens..

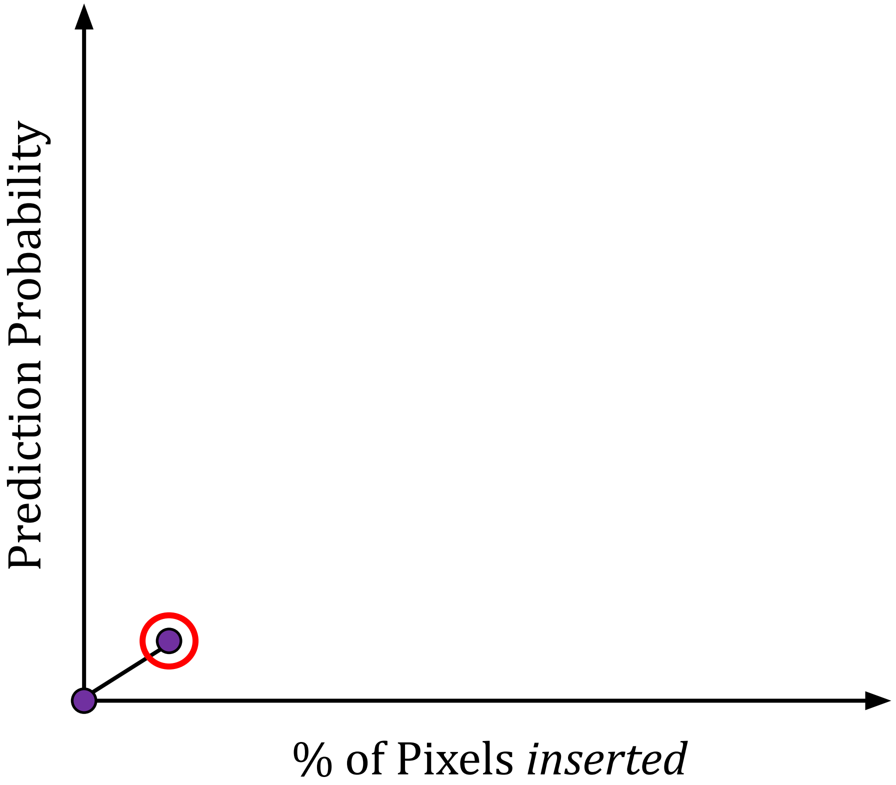




How important are selected features?

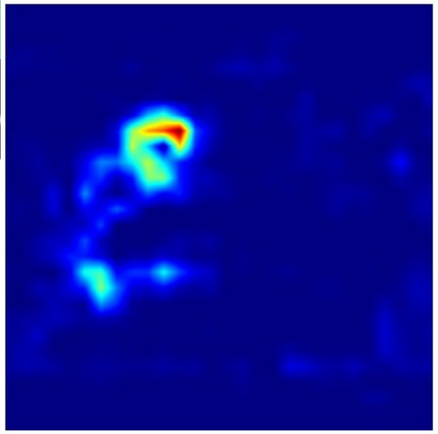
- Insertion: add important features and see what happens..

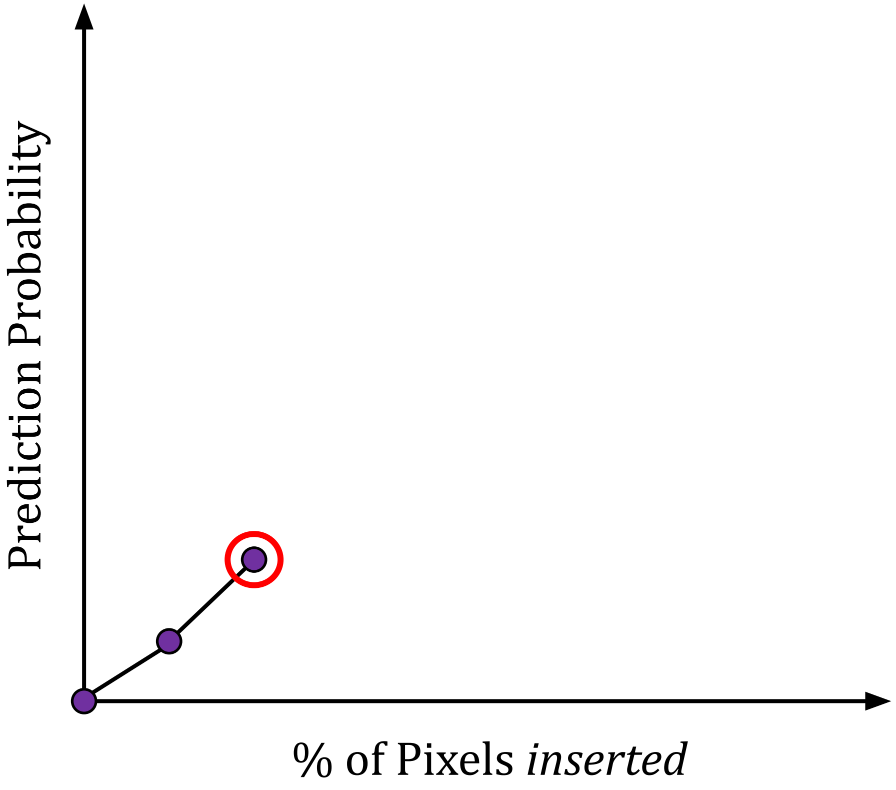




How important are selected features?

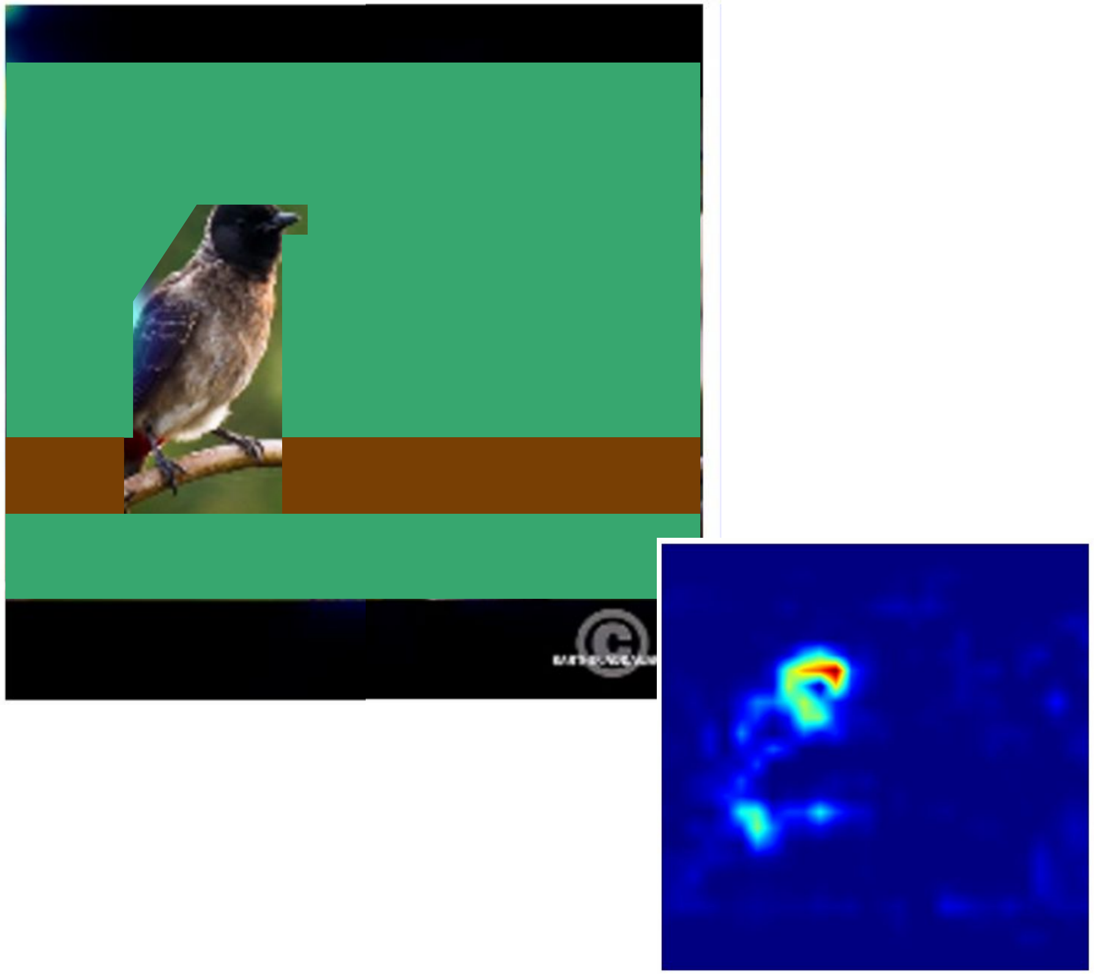
- Insertion: add important features and see what happens..

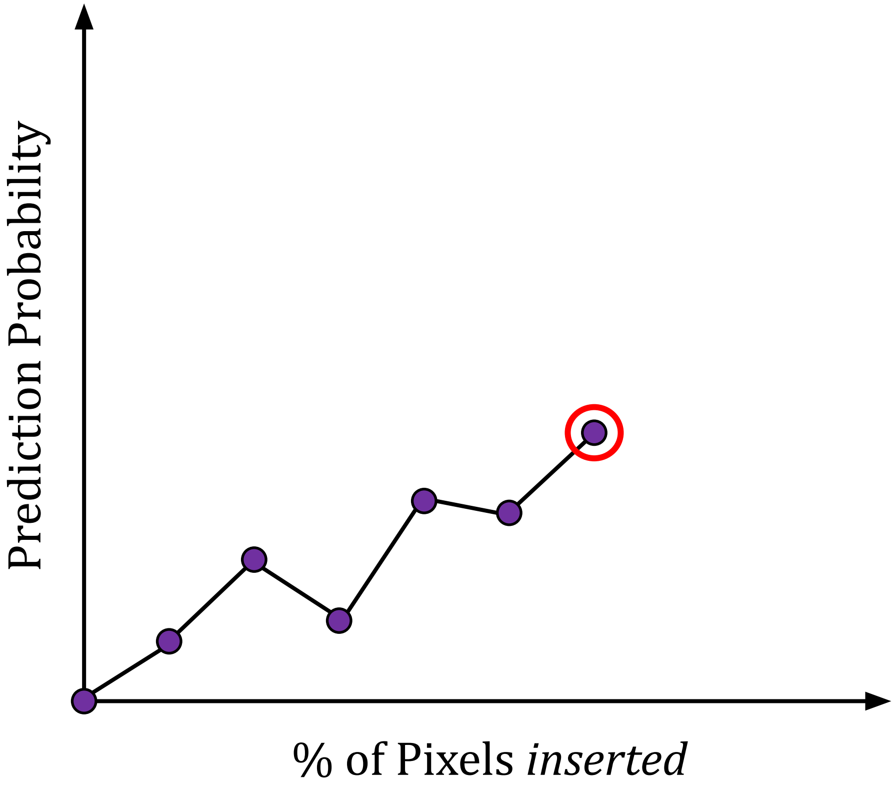




How important are selected features?

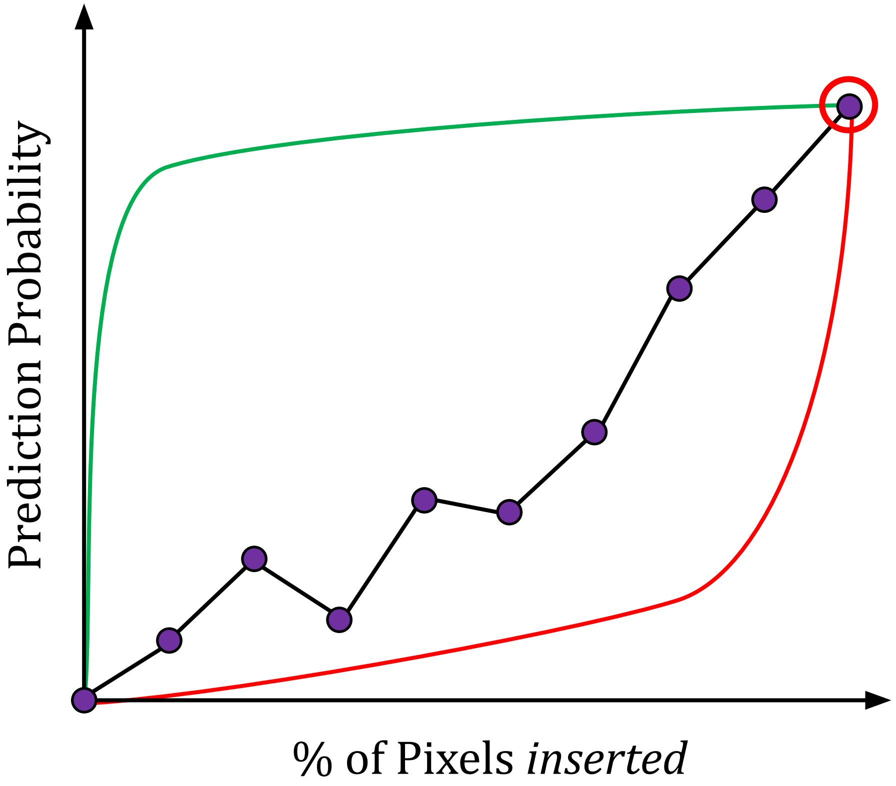
- Insertion: add important features and see what happens..





How important are selected features?

- Insertion: add important features and see what happens..



[Alvarez-Melis, 2018; Agarwal et. al., 2022]

Evaluating Stability of Post hoc Explanations

- Are post hoc explanations unstable w.r.t. small input perturbations?

Local Lipschitz Constant

Post hoc Explanation

$$\hat{L}(x_i) = \max_{x_j \in B_\epsilon(x_i)} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

Input

[Agarwal et. al., 2022]

Evaluating Stability of Post hoc Explanations

- What if the underlying model itself is unstable?
- Relative Output Stability: Denominator accounts for changes in the prediction probabilities
- Relative Representation Stability: Denominator accounts for changes in the intermediate representations of the underlying model

[Dai et. al., 2022]

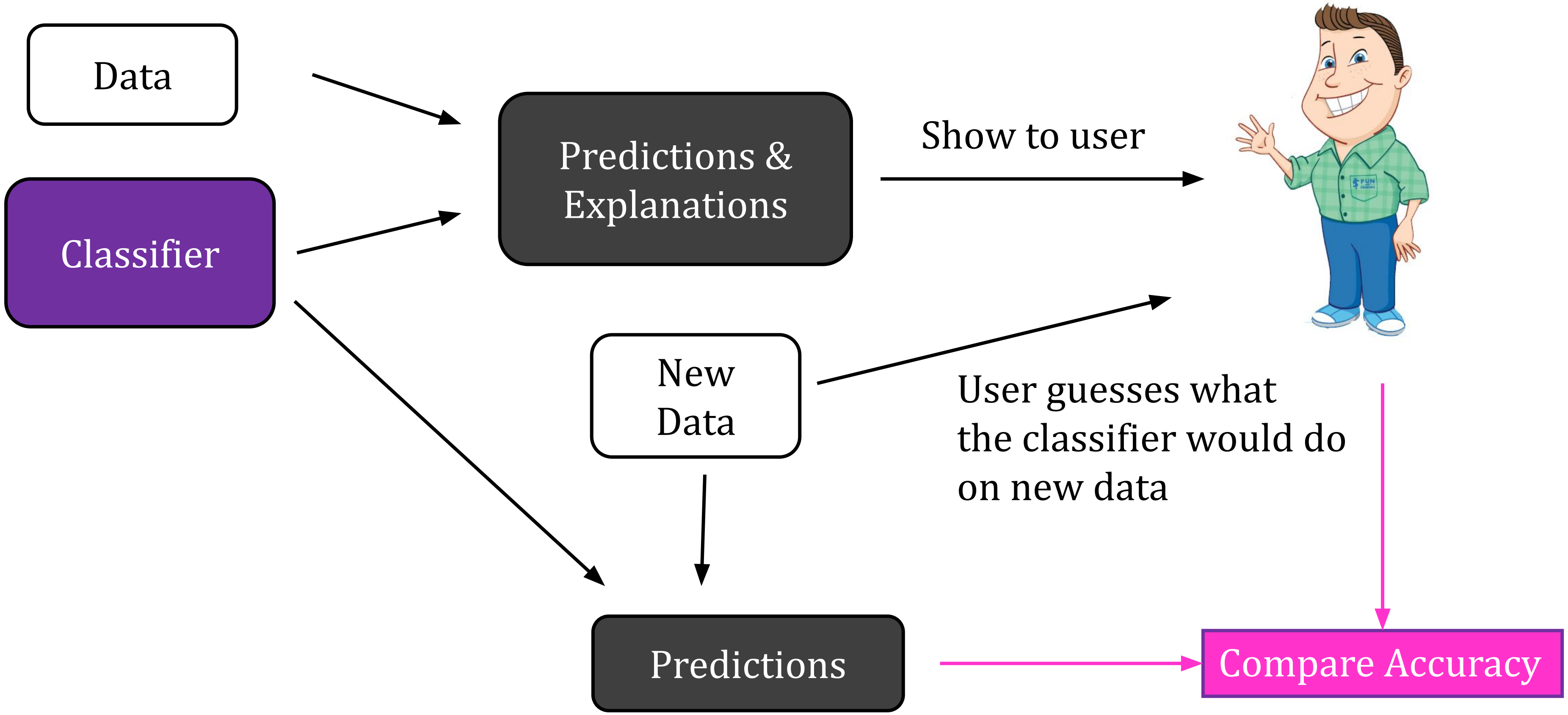
Evaluating Fairness of Post hoc Explanations

- Compute mean faithfulness/stability metrics for instances from majority and minority groups (e.g., race A vs. race B, male vs. female)
- If the difference between the two means is statistically significant, then there is unfairness in the post hoc explanations
- Why/when can such unfairness occur?

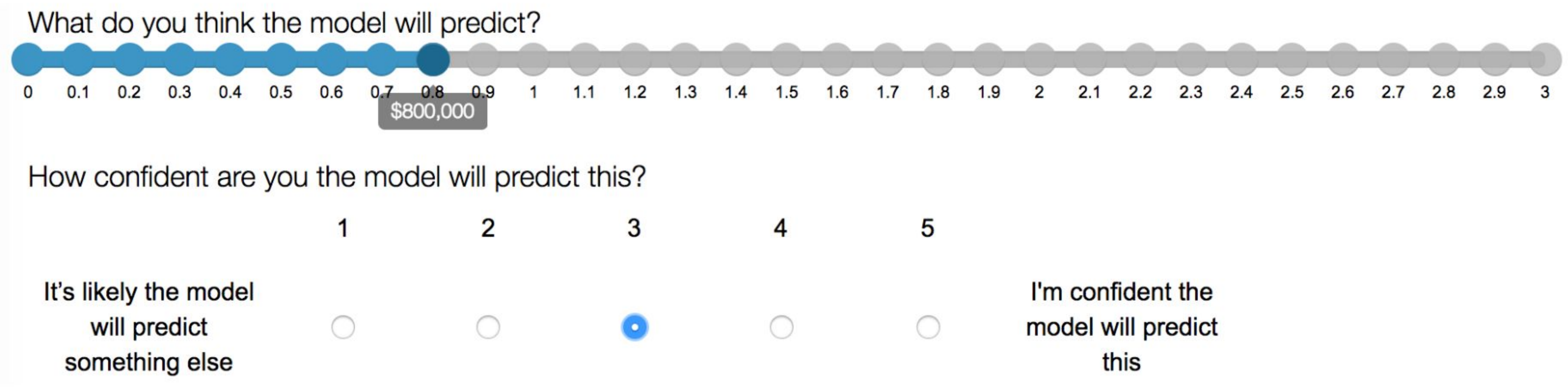
Evaluating Interpretability of Post hoc Explanations

[Ribeiro et al. 2018, Hase and Bansal 2020]

Predicting Behavior ("Simulation")



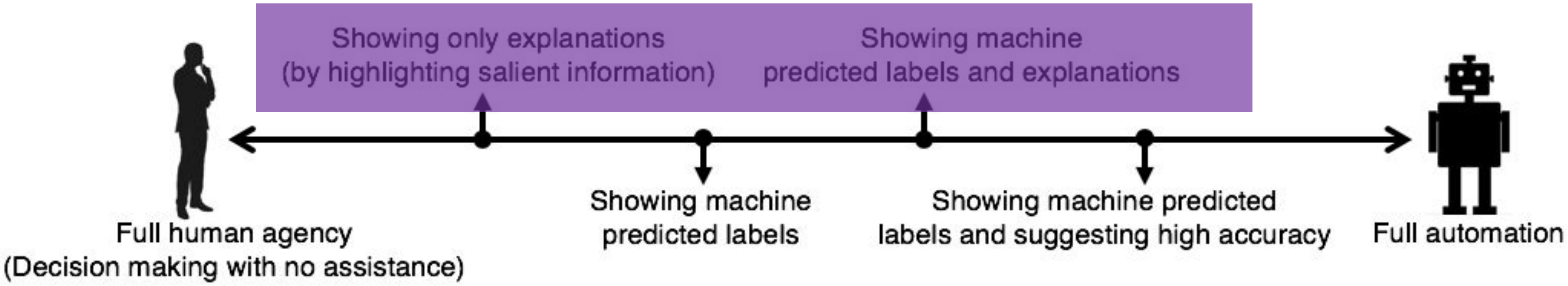
Predicting Behavior ("Simulation")



(a) Step 1: Participants were asked to guess the model's prediction and state their confidence.

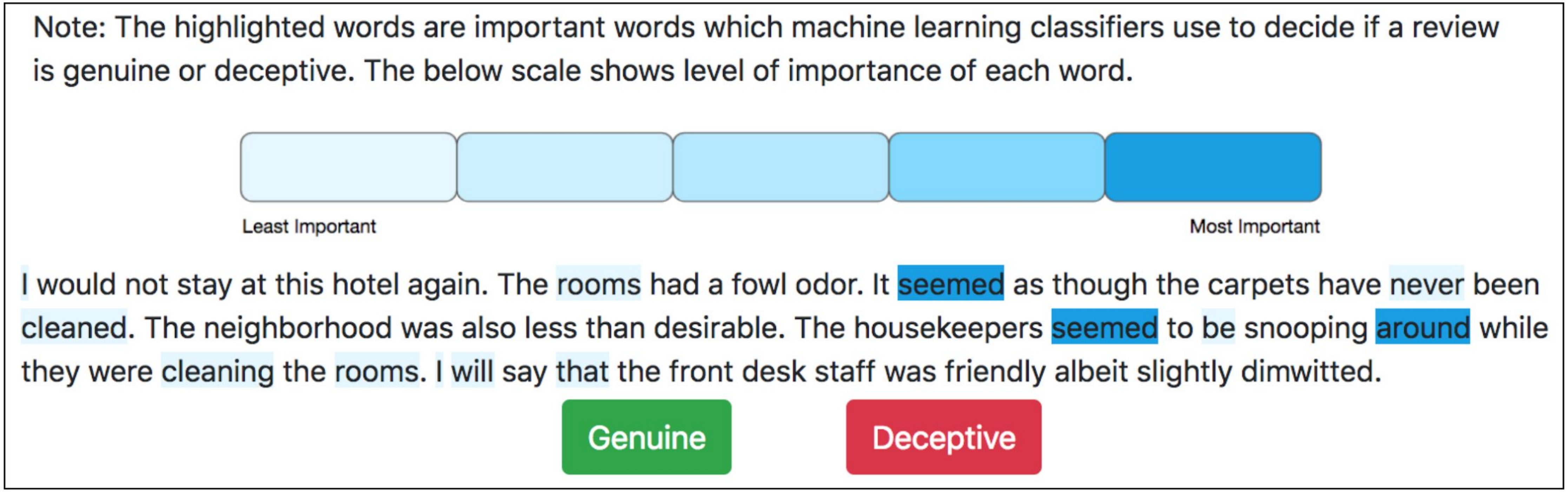
Human-AI Collaboration

- Are Explanations Useful for Making Decisions?
  - For tasks where the algorithms are not reliable by themselves



Human-AI Collaboration

- Deception Detection: Identify fake reviews online
  - Are Humans better detectors with explanations?



https://machineintheloop.com/deception/

Can we improve the accuracy of decisions using feature attribution-based explanations?

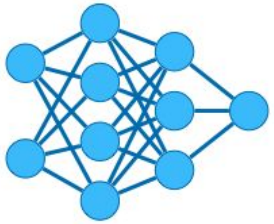
- Prediction Problem: Is a given patient likely to be diagnosed with breast cancer within 2 years?
- User studies carried out with about 78 doctors (Residents, Internal Medicine)
- Each doctor looks at 10 patient records from historical data and makes predictions for each of them.

Lakkaraju et. al., 2022

Can we improve the accuracy of decisions using feature attribution-based explanations?

Accuracy: 78.32%

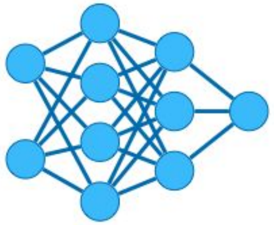
+



At Risk (0.91)

Accuracy: 82.02%

+



At Risk (0.91)

Important Features

ESR
Family Risk
Chronic Health Conditions

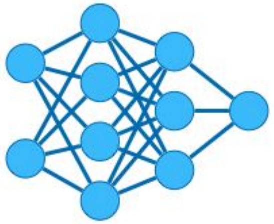
Accuracy: 93.11%

Model Accuracy: 88.92%

Can we improve the accuracy of decisions using feature attribution-based explanations?

Accuracy: 78.32%

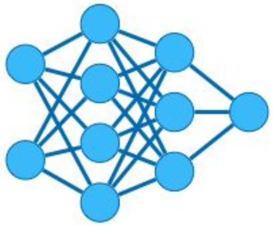
+



At Risk (0.91)

Accuracy: 82.02%

+



At Risk (0.91)

Important Features

Appointment time
Appointment day
Zip code
Doctor ID > 150

Accuracy: 93.11%

Model Accuracy: 88.92%

Challenges of Evaluating Interpretable Models/Post hoc Explanation Methods

- Evaluating interpretations/explanations still an ongoing endeavor
- Parameter settings heavily influence the resulting interpretations/explanations
- Diversity of explanation/interpretation methods diverse metrics
- User studies are not consistent
  - Affected by choice of: UI, phrasing, visualization, population, incentives, …
- All the above leading to conflicting findings

Open Source Tools for Quantitative Evaluation

- Interpretable models: https://github.com/interpretml/interpret
- Post hoc explanation methods: OpenXAI: https://open-xai.github.io/ -- 22 metrics (faithfulness, stability, fairness); public dashboards comparing various metrics on different metrics; 11 lines of code to evaluate explanation quality
- Other XAI libraries: Captum, quantus, shap bench, ERASER (NLP)

Agenda

- ❑ Inherently Interpretable Models
- ❑ Post hoc Explanation Methods
- ❑ Evaluating Model Interpretations/Explanations
- ❑ Empirically & Theoretically Analyzing Interpretations/Explanations
- ❑ Future of Model Understanding

Empirically Analyzing Interpretations/Explanations

- Lot of recent focus on analyzing the behavior of post hoc explanation methods.

- Empirical studies analyzing the faithfulness, stability, fairness, adversarial vulnerabilities, and utility of post hoc explanation methods.

- Several studies demonstrate limitations of existing post hoc methods.

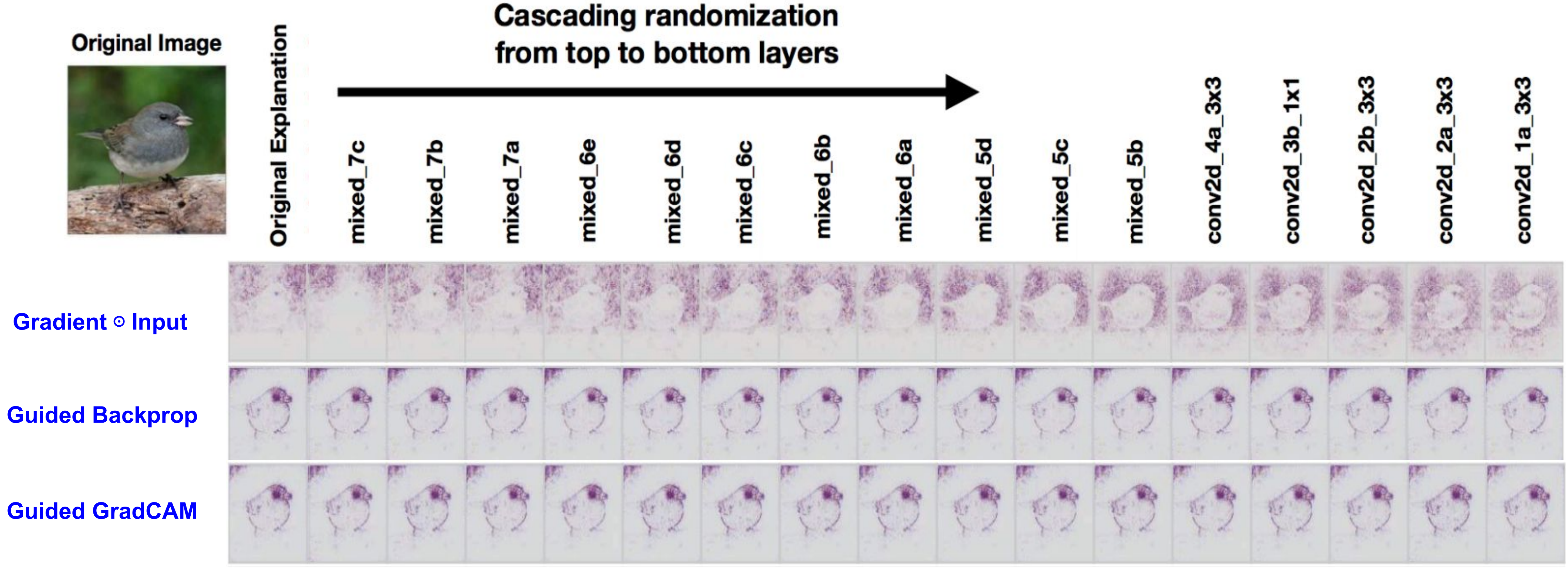
Limitations: **Faithfulness**

Model parameter randomization test

[Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.](#)

Limitations: **Faithfulness**

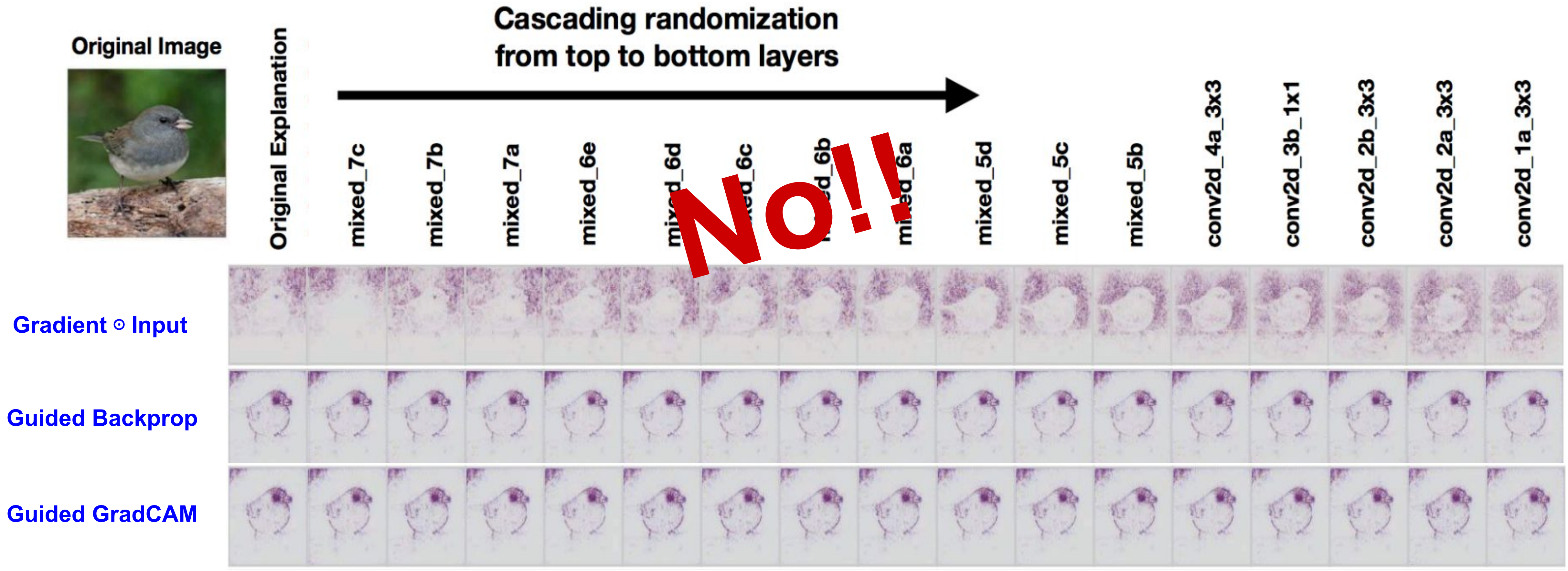
Model parameter randomization test



[Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.](#)

Limitations: **Faithfulness**

Model parameter randomization test



[Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.]

Limitations: **Faithfulness**

Randomizing class labels of instances also didn't impact explanations!

[Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.]

Limitations: **Stability**

Are post-hoc explanations unstable wrt small non-adversarial input perturbation?

Local Lipschitz Constant

Explanation function: LIME, SHAP, Gradient...etc.

$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

Input

input model

$$A(x, f, H)$$

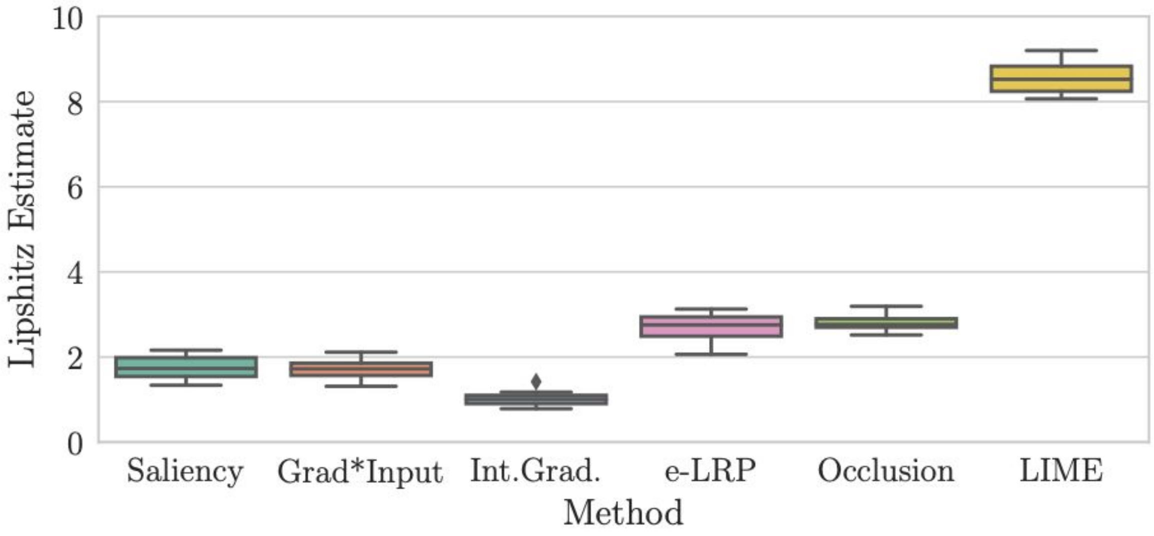
hyperparameters

[Alvarez-Melis, David et al. "On the robustness of interpretability methods." WHI, ICML, 2018.](#)

Limitations: **Stability**

Are post-hoc explanations unstable wrt small non-adversarial input perturbation?

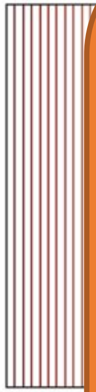
- Perturbation approaches like LIME can be unstable.



Estimate for 100 tests for an MNIST Model.

[Alvarez-Melis, David et al. "On the robustness of interpretability methods." WHI, ICML, 2018.]

Limitations: Stability – Problem is Worse!



(a

Problem with having too few perturbations?
If so, what is the optimal number of perturbations?

When you repeatedly run LIME on the same instance, you get different explanations (blue region)

Post-hoc Explanations are Fragile

Post-hoc explanations can be easily manipulated.

Original Image

[Dombrowski et. al. 2019]

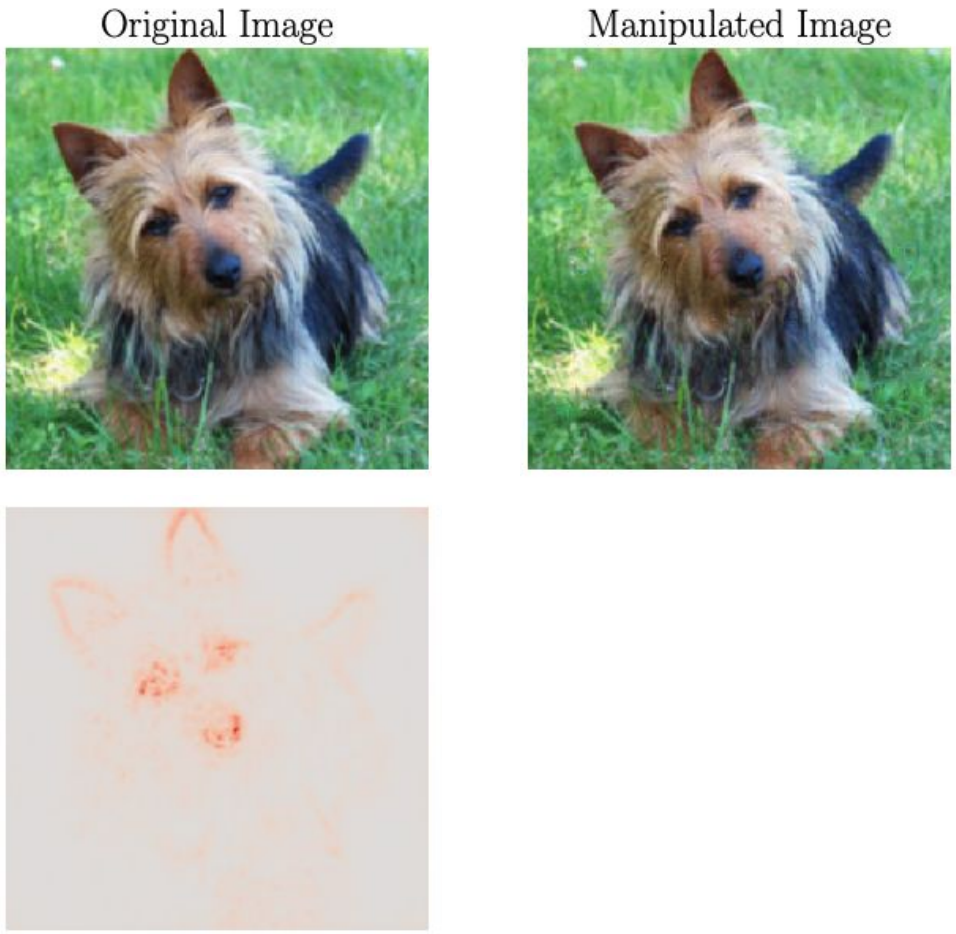
Post-hoc Explanations are Fragile

Post-hoc explanations can be easily manipulated.

[Dombrowski et. al. 2019]

Post-hoc Explanations are Fragile

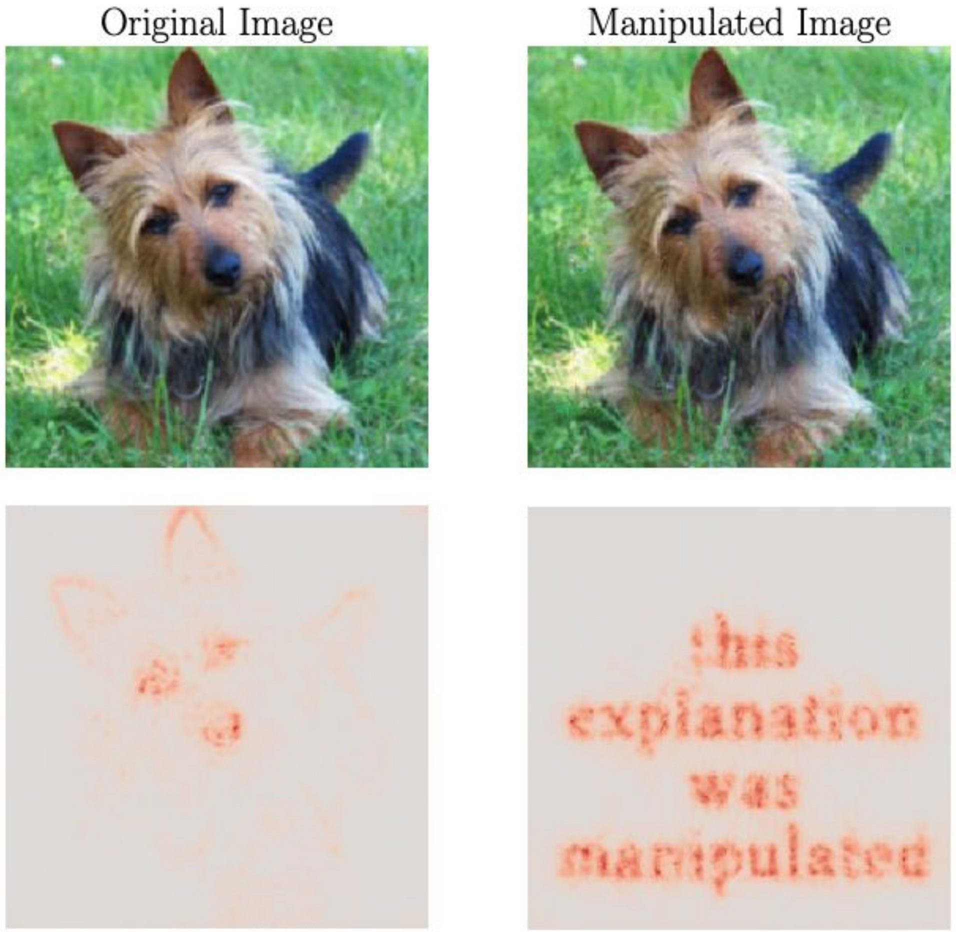
Post-hoc explanations can be easily manipulated.



[Dombrowski et. al. 2019]

Post-hoc Explanations are Fragile

Post-hoc explanations can be easily manipulated.



Dombrowski et. al. 2019

Adversarial Attacks on Explanations

Minimally modify the input with a **small perturbation without changing the model prediction**.

$$\arg\max_{\boldsymbol{\delta}} \mathcal{D}\left(\boldsymbol{I}(\boldsymbol{x}_t;\mathcal{N}),\boldsymbol{I}(\boldsymbol{x}_t+\boldsymbol{\delta};\mathcal{N})\right)$$

Adversarial Attack of Ghorbani et. al. 2018

Adversarial Attacks on Explanations

Minimally modify the input with a **small perturbation without changing the model prediction**.

$$\arg\max_{\boldsymbol{\delta}} \mathcal{D}\left(\boldsymbol{I}(\boldsymbol{x}_t;\mathcal{N}), \boldsymbol{I}(\boldsymbol{x}_t+\boldsymbol{\delta};\mathcal{N})\right)$$

$$\text{subject to: } ||\boldsymbol{\delta}||_\infty \leq \epsilon,$$

Adversarial Attack of Ghorbani et. al. 2018

Adversarial Attacks on Explanations

Minimally modify the input with a **small perturbation without changing the model prediction**.

$$\arg\max_{\boldsymbol{\delta}} \mathcal{D}\left(\boldsymbol{I}(\boldsymbol{x}_t;\mathcal{N}),\boldsymbol{I}(\boldsymbol{x}_t+\boldsymbol{\delta};\mathcal{N})\right)$$

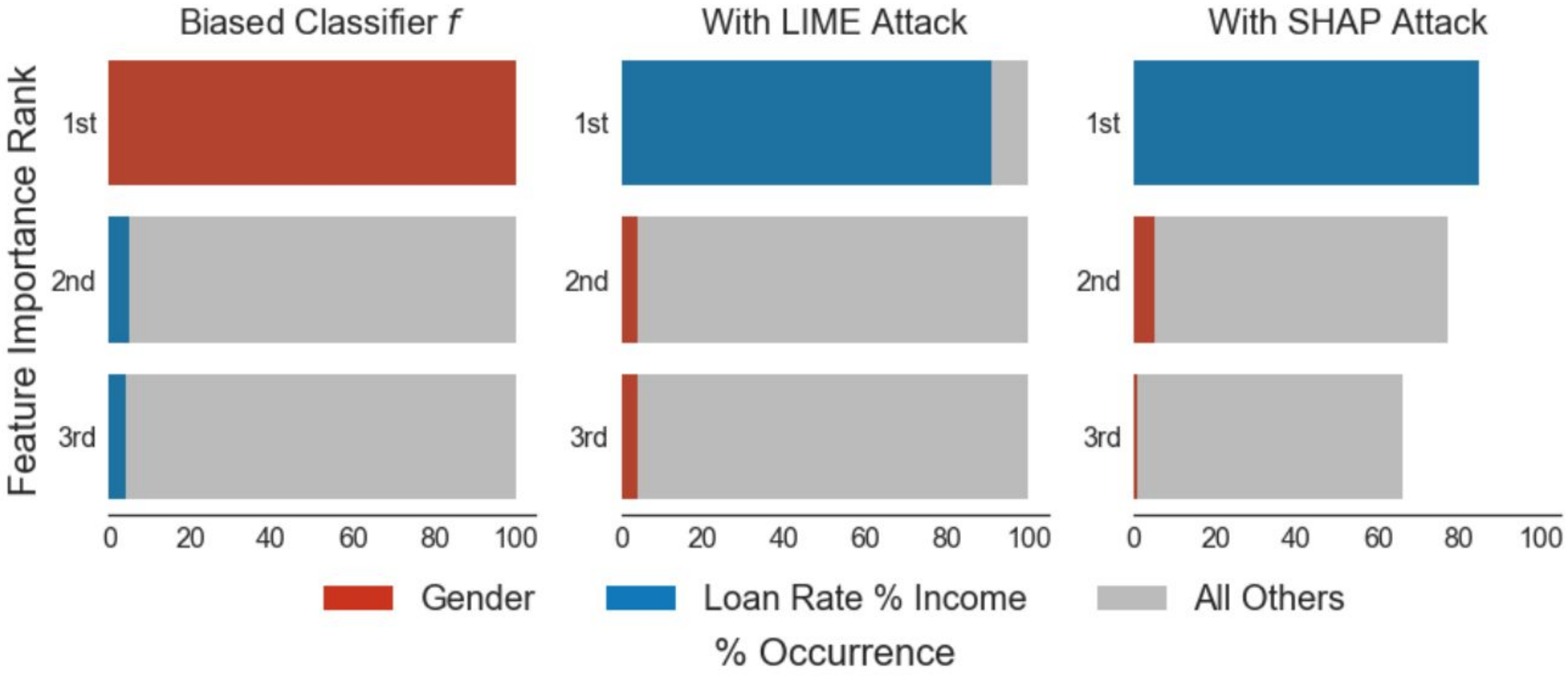
$$\text{subject to: } ||\boldsymbol{\delta}||_\infty \leq \epsilon,$$

$$\text{Prediction}(\boldsymbol{x}_t+\boldsymbol{\delta};\mathcal{N}) = \text{Prediction}(\boldsymbol{x}_t;\mathcal{N})$$

Adversarial Attack of Ghorbani et. al. 2018

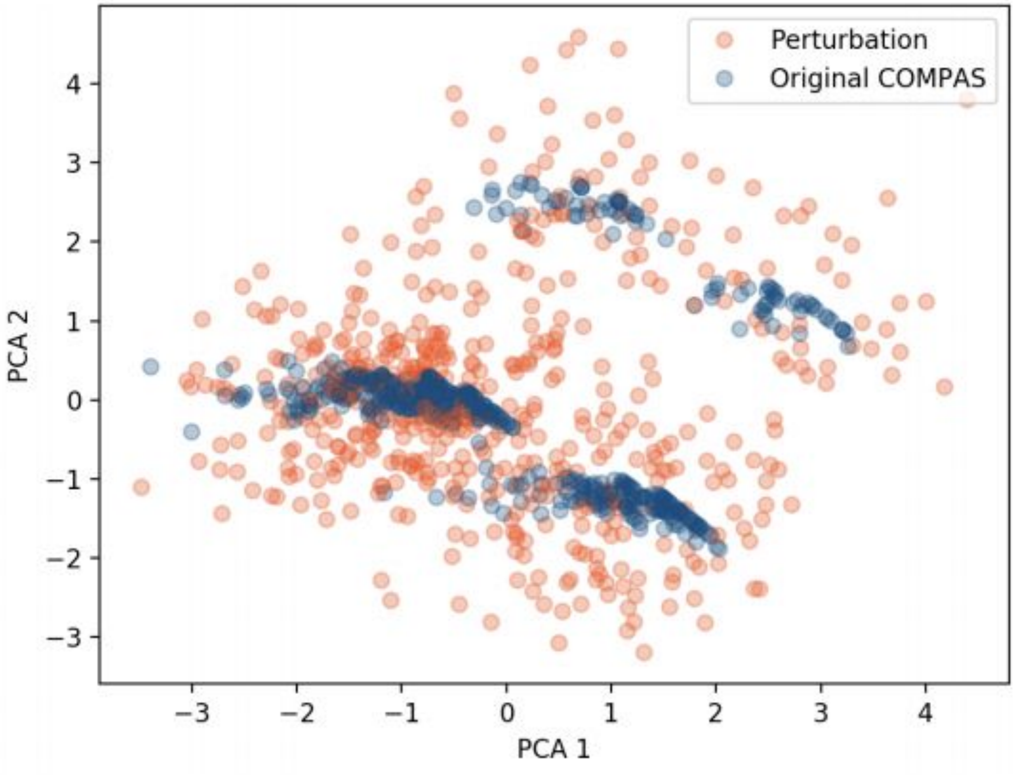
Adversarial Classifiers to fool LIME & SHAP

Scaffolding attack used to **hide classifier dependence on gender**.



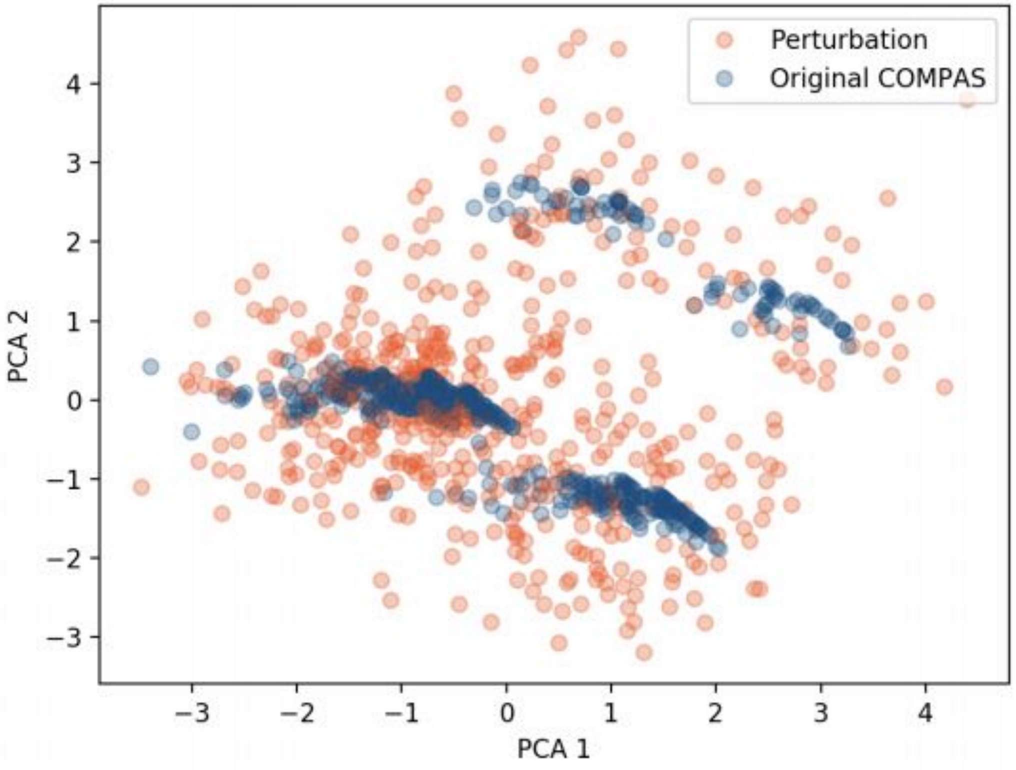
[Slack and Hilgard et. al. 2020]

Vulnerabilities of LIME/SHAP: Intuition



Several perturbed data points are out of distribution (OOD)!

Vulnerabilities of LIME/SHAP: Intuition



Adversaries can exploit this and build a classifier that is biased on in-sample data points and unbiased on OOD samples!

Building Adversarial Classifiers

- **Setting:**
  - Adversary wants to deploy a biased classifier f in real world.
    - E.g., uses only race to make decisions
  - Adversary must provide black box access to customers and regulators who may use post hoc techniques (GDPR).
  - *Goal of adversary is to fool post hoc explanation techniques and hide underlying biases of f*

Building Adversarial Classifiers

- **Input:** Adversary provides us with the biased classifier *f*, an input dataset $\boldsymbol{X}$ sampled from real world input distribution $\boldsymbol{X}_{\text{dist}}$

- **Output:** Scaffolded classifier *e* which behaves exactly like f when making predictions on instances sampled from $\boldsymbol{X}_{\text{dist}}$ but will not reveal underlying biases of f when probed with perturbation-based post hoc explanation techniques.
  - *e* is the adversarial classifier

Building Adversarial Classifiers

- Adversarial classifier e can be defined as:

$$e(x) = \begin{cases} f(x), & \text{if } x \in \mathcal{X}_{dist} \\ \psi(x), & \text{otherwise} \end{cases}$$

- f is the biased classifier input by adversary.

- ψ is the unbiased classifier (e.g., only uses features uncorrelated to sensitive attributes)

Limitations: Stability

Post-hoc explanations can be unstable to small, **non-adversarial**, perturbations to the input.

Alvarez et. al. 2018.

Limitations: Stability

Post-hoc explanations can be unstable to small, **non-adversarial**, perturbations to the input.

'Local Lipschitz Constant'

Explanation function: LIME, SHAP, Gradient…etc.

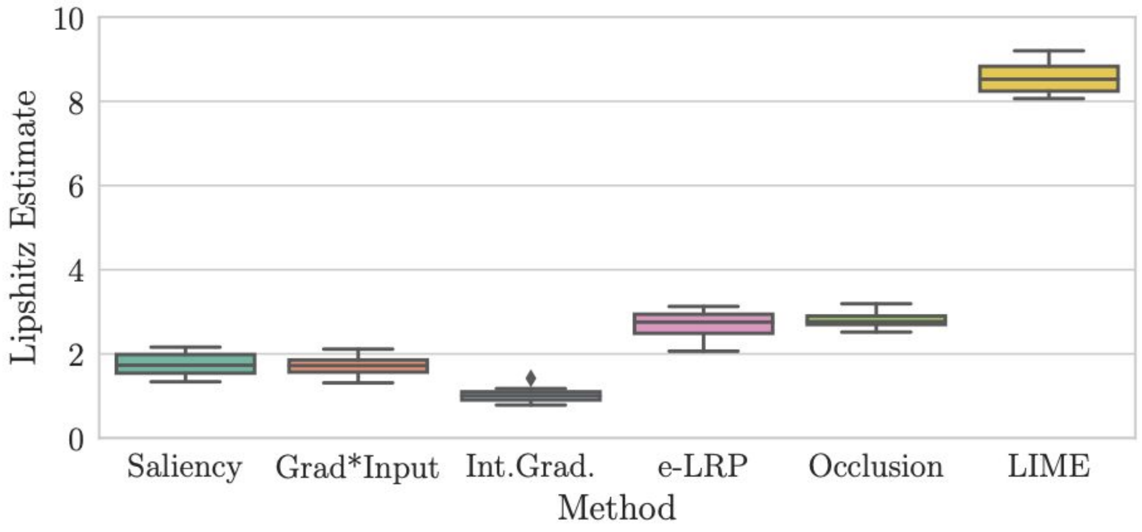
$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

Input

Alvarez et. al. 2018.

Limitations: Stability

- Perturbation approaches like LIME can be unstable.

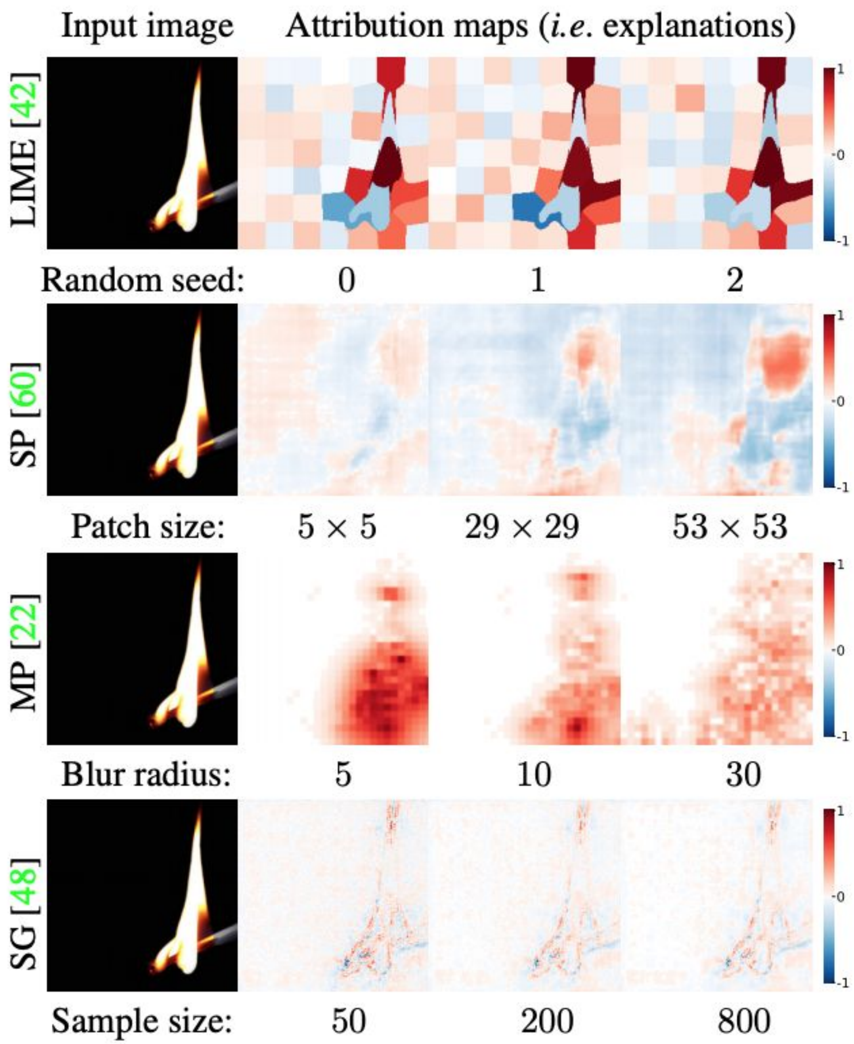


Estimate for 100 tests for an MNIST Model.

[Alvarez et. al. 2018.]

Sensitivity to Hyperparameters

Explanations can be highly sensitive to hyperparameters such as **random seed**, number of perturbations, patch size, etc.



[Bansal, Agarwal, & Nguyen, 2020.](#)

Utility: High fidelity explanations can mislead

In a bail adjudication task, **misleading** high-fidelity explanations improve end-user (domain experts) trust.

True Classifier relies on race

If Race ≠ African American:
- If Prior-Felony = Yes and Crime-Status = Active, then Risky
- If Prior-Convictions = 0, then Not Risky

If Race = African American:
- If Pays-rent = No and Gender = Male, then Risky
- If Lives-with-Partner = No and College = No, then Risky
- If Age ≥35 and Has-Kids = Yes, then Not Risky
- If Wages ≥70K, then Not Risky

Default: Not Risky

Lakkaraju & Bastani 2019.

Utility: High fidelity explanations can mislead

In a bail adjudication task, **misleading** high-fidelity explanations improve end-user (domain experts) trust.

True Classifier relies on race

If Race ≠ African American:
- If Prior-Felony = Yes and Crime-Status = Active, then Risky
- If Prior-Convictions = 0, then Not Risky

If Race = African American:
- If Pays-rent = No and Gender = Male, then Risky
- If Lives-with-Partner = No and College = No, then Risky
- If Age ≥35 and Has-Kids = Yes, then Not Risky
- If Wages ≥70K, then Not Risky

Default: Not Risky

High fidelity 'misleading' explanation

If Current-Offense = Felony:
- If Prior-FTA = Yes and Prior-Arrests ≥ 1, then Risky
- If Crime-Status = Active and Owns-House = No and Has-Kids = No, then Risky
- If Prior-Convictions = 0 and College = Yes and Owns-House = Yes, then Not Risky

If Current-Offense = Misdemeanor and Prior-Arrests > 1:
- If Prior-Jail-Incarcerations = Yes, then Risky
- If Has-Kids = Yes and Married = Yes and Owns-House = Yes, then Not Risky
- If Lives-with-Partner = Yes and College = Yes and Pays-Rent = Yes, then Not Risky

If Current-Offense = Misdemeanor and Prior-Arrests ≤ 1:
- If Has-Kids = No and Owns-House = No and Prior-Jail-Incarcerations = Yes, then Risky
- If Age ≥ 50 and Has-Kids = Yes and Prior-FTA = No, then Not Risky

Default: Not Risky

Lakkaraju & Bastani 2019.

Utility: Post hoc Explanations Instill Over Trust

- Domain experts and end users seem to be over trusting explanations & the underlying models based on explanations

    - Data scientists over trusted explanations without even comprehending them -- "*Participants trusted the tools because of their visualizations and their public availability*"

Responses from Data Scientists Using Explainability Tools (GAM and SHAP)

"I didn't fully grasp what SHAP values were. This is a pretty popular tool and I get the log-odds concept in general. I figure they were showing SHAP values for a reason. Maybe it's easier to judge relationships using log-odds instead of predicted value. Anyway, so it made sense I suppose." (P6, SHAP)

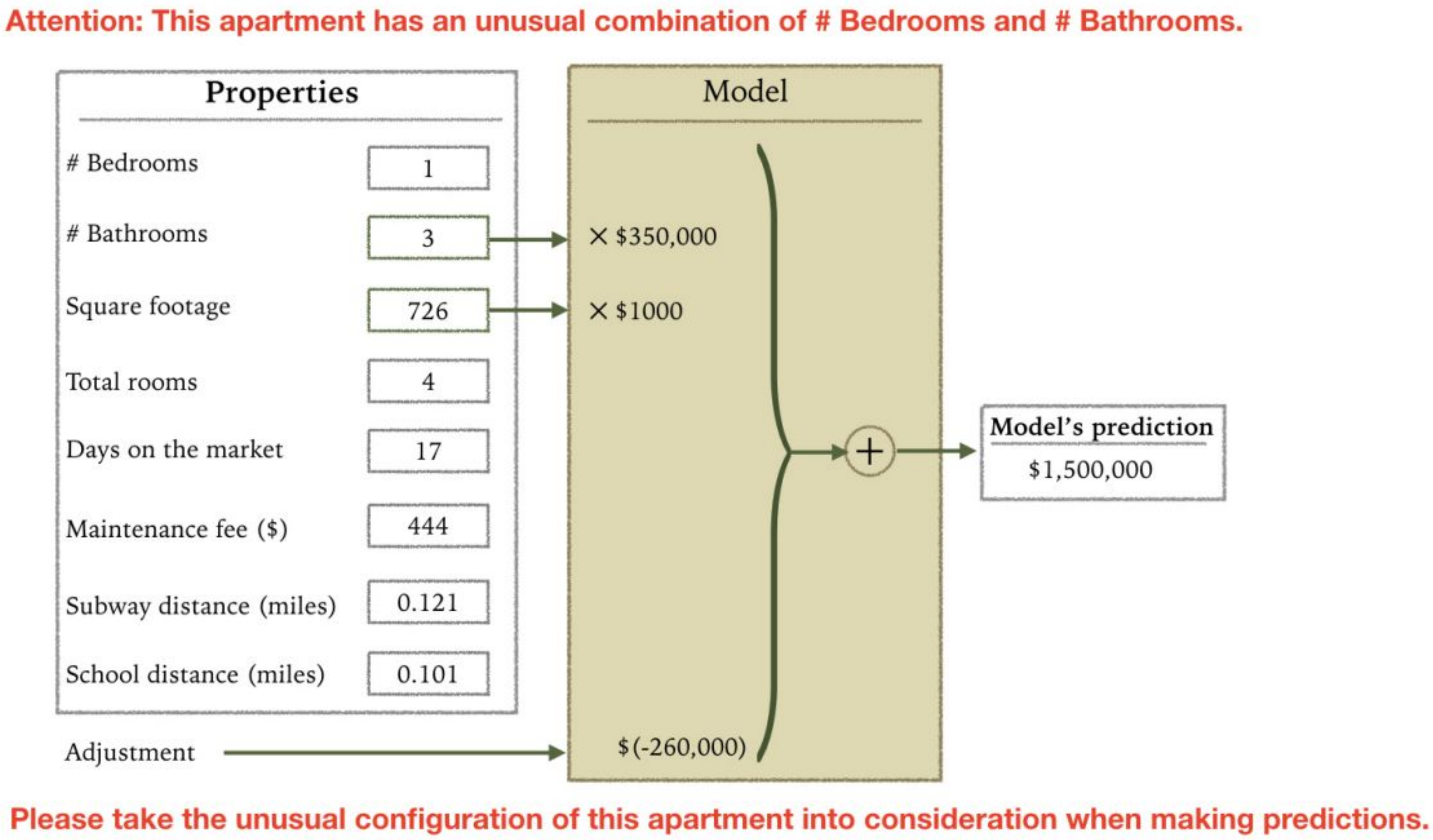
"Age 38 seems to have the highest positive influence on income based on the plot. Not sure why, but the explanation clearly shows it... makes sense." (P9, GAMs)

"[The tool] assigns a value that is important to know, but it's showing that in a way that makes you misinterpret that value. Now I want to go back and check all my answers"... [later] "Okay, so, it's not showing me a whole lot more than what I can infer on my own. Now I'm thinking... is this an 'interpretability tool'?" (P4, SHAP)

"[The tool] shows visualizations of ML models, which is not something anything else I have worked with has done. It's very transparent, and that makes me trust it more" (P9, GAMs).

Utility: Explanations for Debugging

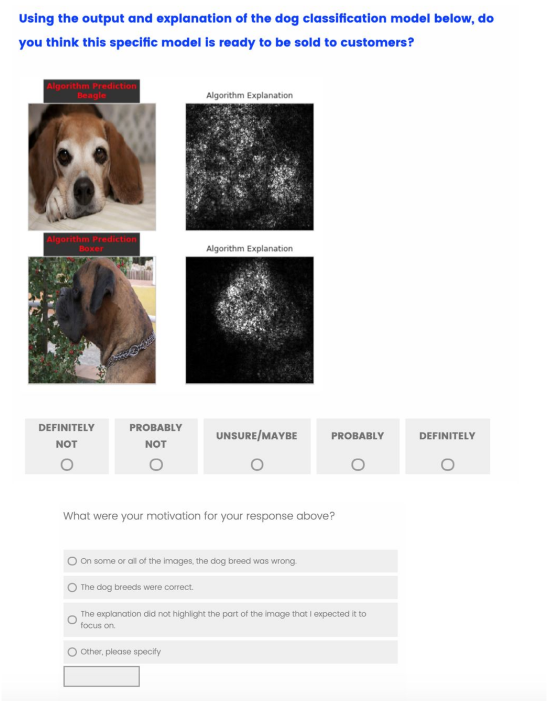
In a housing price prediction task, Amazon mechanical turkers are unable to use linear model coefficients to diagnose model mistakes.



[Poursabzi-Sangdeh et. al. 2019]()

Utility: Explanations for Debugging

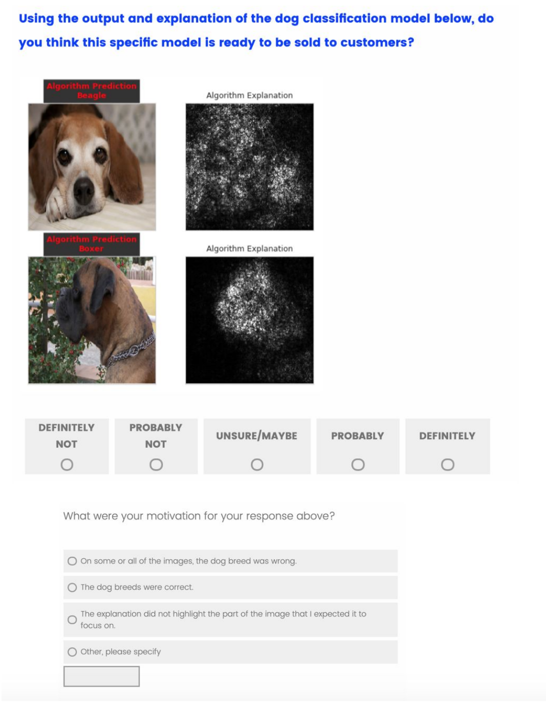
In a dog breeds classification task, users familiar with machine learning **rely on labels, instead of saliency maps,** for diagnosing model errors.

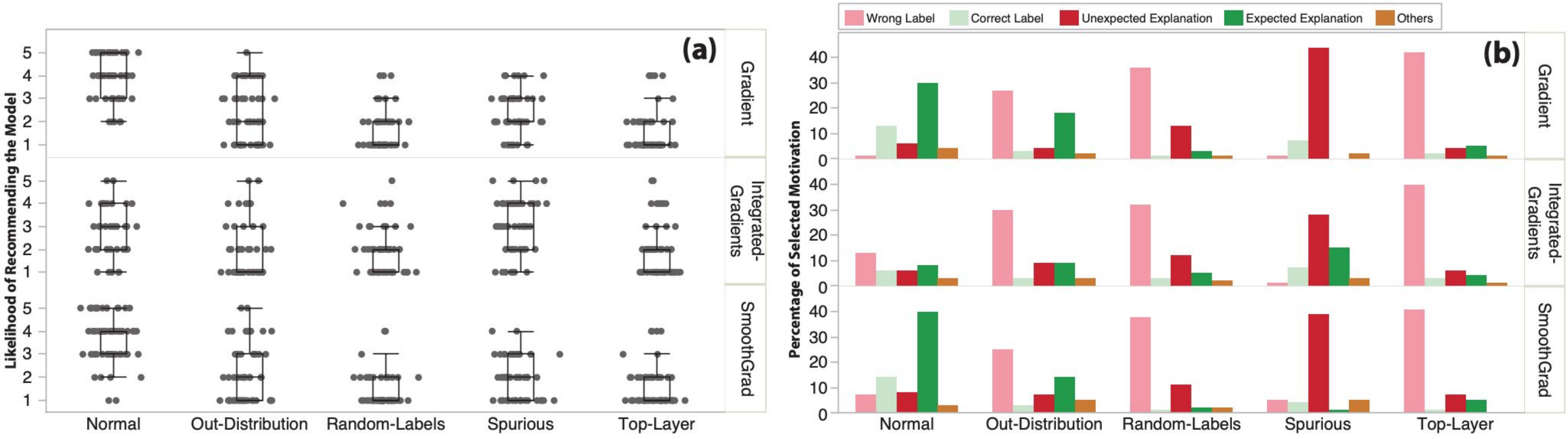


[Adebayo et. al., 2020.]

Utility: Explanations for Debugging

In a dog breeds classification task, users familiar with machine learning **rely on labels, instead of saliency maps,** for diagnosing model errors.





[Adebayo et. al., 2020.]()

Conflicting Evidence on Utility of Explanations

- **Mixed evidence**:
    - simulation and benchmark studies show that explanations are useful for debugging;
    - however, recent user studies show limited utility in practice.

Conflicting Evidence on Utility of Explanations

- **Mixed evidence**:
    - simulation and benchmark studies show that explanations are useful for debugging;
    - however, recent user studies show limited utility in practice.

- Rigorous **user studies** and **pilots with end-users** can continue to help provide feedback to researchers on what to address (see: Alqaraawi et. al. 2020, Bhatt et. al. 2020 & Kaur et. al. 2020).

Utility: Disagreement Problem in XAI

- Study to understand:
  - if and how often feature attribution based explanation methods disagree with each other in practice
  - What constitutes disagreement between these explanations, and how to formalize the notion of explanation disagreement based on practitioner inputs?
  - How do practitioners resolve explanation disagreement?

Krishna and Han et. al., 2022

Practitioner Inputs on Explanation Disagreement

- 30 minute semi-structured interviews with 25 data scientists
- 84% of participants said they often encountered disagreement between explanation methods
- Characterizing disagreement:
  - Top features are different
  - Ordering among top features is different
  - Direction of top feature contributions is different
  - Relative ordering of features of interest is different

How do Practitioners Resolve Disagreements?

- Online user study where 25 users were shown explanations that disagree and asked to make a choice, and explain why
- Practitioners are choosing methods due to:
  - Associated theory or publication time (33%)
  - Explanations matching human intuition better (32%)
  - Type of data (23%)
    - E.g., LIME or SHAP are better for tabular data

How do Practitioners Resolve Disagreements?

Algorithm	Reasons that algorithm was chosen in disagreement
KernelSHAP	• [36%] SHAP is better for tabular data (*"SHAP is more commonly used [than Gradient] for tabular data"*) • [25%] SHAP is more familiar (*"More information present + more familiarity"*) • [14%] SHAP is a better algorithm overall (*"SHAP seems more methodical than LIME"*, *"SHAP is a more rigorous approach [than LIME] in theory"*)
SmoothGrad	• [33%] SmoothGrad paper is newer or better (*"SmoothGrad is apparently more robust"*, *"SmoothGrad is often considered improved verison of grad"*) • [58%] Reasons based on the explainability map shown (*"directionality of the attributions ... [agree] with intuition"*, *"gradient has unstability problems [, so] smoothgrad"*)
LIME	• [54%] LIME is better for tabular data (*"I use LIME for structured data."*) • [15%] LIME is more familiar/easier to interpret (*"I am more familiar with LIME"*, *"LIME is easy to interpret"*)
Integrated Gradients	• [86%] Integrated Gradients paper is better (*"IG came after gradients and paper shows improvements"*, *"integrated gradients paper showed improvements [over Gradient × Input]"*)

Empirical Analysis: Summary

- **Faithfulness/Fidelity**
  - Some explanation methods do not '*reflect*' the underlying model.

- **Fragility**
  - Post-hoc explanations can be easily manipulated.

- **Stability**
  - Slight changes to inputs can cause large changes in explanations.

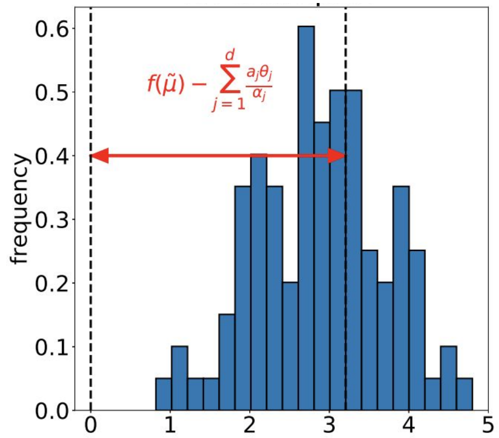
- **Useful in practice?**

Theoretically Analyzing Interpretable Models

- Two main classes of theoretical results:
- Interpretable models learned using certain algorithms are certifiably optimal
  - E.g., rule lists (Angelino et. al., 2018)
- No accuracy-interpretability tradeoffs in certain settings
  - E.g., reinforcement learning for mazes (Mansour et. al., 2022)

Theoretical Analysis of Tabular LIME w.r.t. Linear Models

- Theoretical analysis of LIME
  - "black box" is a *linear model*
  - data is *tabular* and *discretized*



- Obtained closed-form solutions of the average coefficients of the "surrogate" model (explanation output by LIME)

- The coefficients obtained are proportional to the gradient of the function to be explained

- Local error of surrogate model is bounded away from zero with high probability

[Agarwal et. al., 2020]

Unification and Robustness of LIME and SmoothGrad

- C-LIME (a continuous variant of LIME) and SmoothGrad converge to the same explanation in expectation
- At expectation, the resulting explanations are provably robust according to the notion of Lipschitz continuity
- Finite sample complexity bounds for the number of perturbed samples required for SmoothGrad and C-LIME to converge to their expected output

Function Approximation Perspective to Characterizing Post hoc Explanation Methods

- Various feature attribution methods (e.g., LIME, C-LIME, KernelSHAP, Occlusion, Vanilla Gradients, Gradient times Input, SmoothGrad, Integrated Gradients) are essentially local linear function approximations.

$$g^* = \arg\min_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{\xi \sim \mathcal{Z}} \ell(f, g, \mathbf{x}_0, \xi)$$

- But…

[Han et. al., 2022]

Function Approximation Perspective to Characterizing Post hoc Explanation Methods

- But, they adopt different loss functions, and local neighborhoods

Explanation Method	**Local Neighborhood $\mathcal{Z}$ around $\mathbf{x}_0$**	**Loss Function ℓ**
C-LIME	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Squared Error
SmoothGrad	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Gradient Matching
Vanilla Gradients	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2), \sigma \to 0$	Gradient Matching
Integrated Gradients	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(0, 1)$	Gradient Matching
Gradients × Input	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(a, 1), a \to 1$	Gradient Matching
LIME	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Exponential kernel	Squared Error
KernelSHAP	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Shapley kernel	Squared Error
Occlusion	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Random one-hot vectors	Squared Error

Function Approximation Perspective to Characterizing Post hoc Explanation Methods

- *No Free Lunch Theorem for Explanation Methods*: No single method can perform optimally across all neighborhoods

Theorem 3 (No Free Lunch for Explanation Methods). *Consider the scenario where we explain a black-box model f around point $\mathbf{x}_0$ using an interpretable model g from class $\mathcal{G}$ and a valid loss function ℓ where the distance between f and $\mathcal{G}$ is given by $\mathrm{d}(f, \mathcal{G}) = \min_{g \in \mathcal{G}} \max_{\mathbf{x} \in \mathcal{X}} \ell(f, g, 0, \mathbf{x})$. Then, for any explanation g^* on a neighborhood distribution $\xi_1 \sim \mathcal{Z}_1$ such that $\max_{\xi_1} \ell(f, g^*, \mathbf{x}_0, \xi_1) \leq \epsilon$, we can always find another neighborhood $\xi_2 \sim \mathcal{Z}_2$ such that $\max_{\xi_2} \ell(f, g^*, \mathbf{x}_0, \xi_2) \geq \mathrm{d}(f, \mathcal{G})$.*

Agenda

- Inherently Interpretable Models
- Post hoc Explanation Methods
- Evaluating Model Interpretations/Explanations
- Empirically & Theoretically Analyzing Interpretations/Explanations
- Future of Model Understanding

Future of Model Understanding

Methods for More Reliable
Post hoc Explanations

Model Understanding
Beyond Classification

Theoretical Analysis of the Behavior of
Interpretable Models & Explanation Methods

Intersections with Model Robustness

Empirical Evaluation of the Correctness &
Utility of Model Interpretations/Explanations

Intersections with Model Fairness

Characterizing Similarities and Differences
Between Various Methods

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

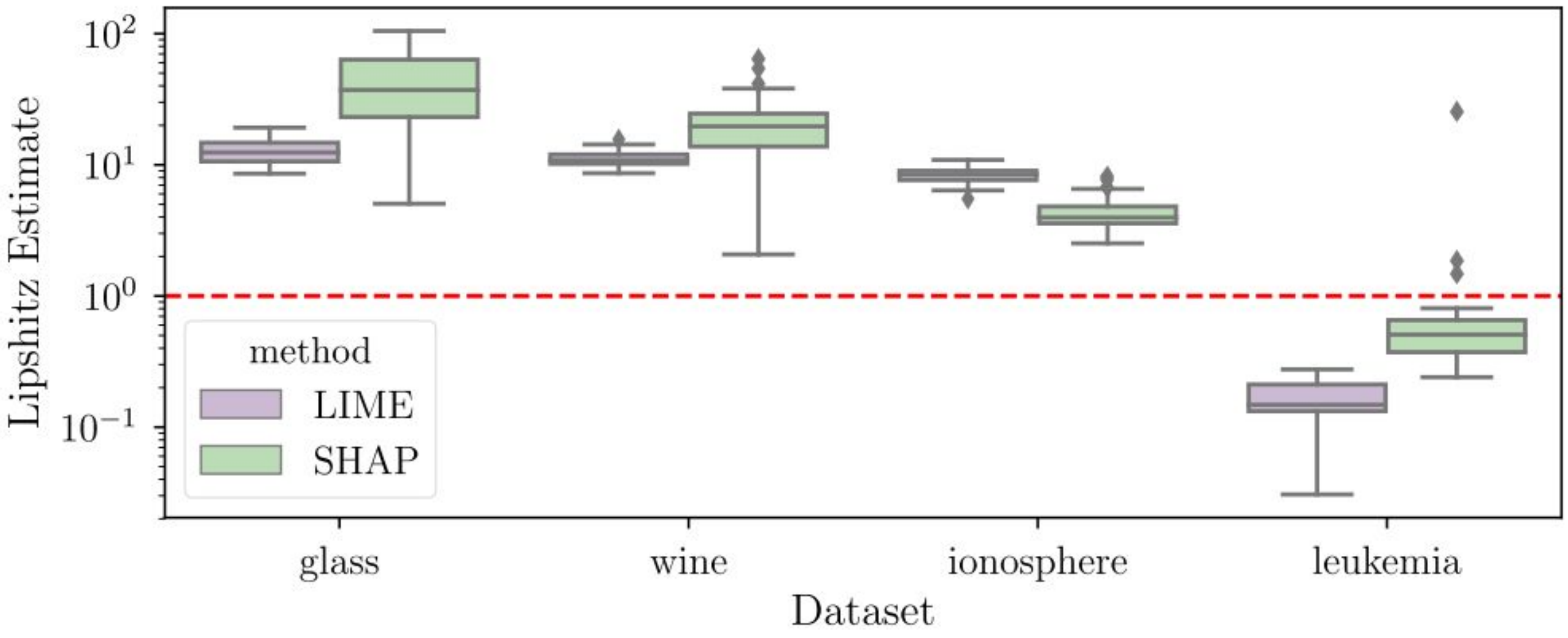
Methods for More Reliable Post hoc Explanations

- We have seen several limitations in the behavior of post hoc explanation methods – e.g., unstable, inconsistent, fragile, not faithful

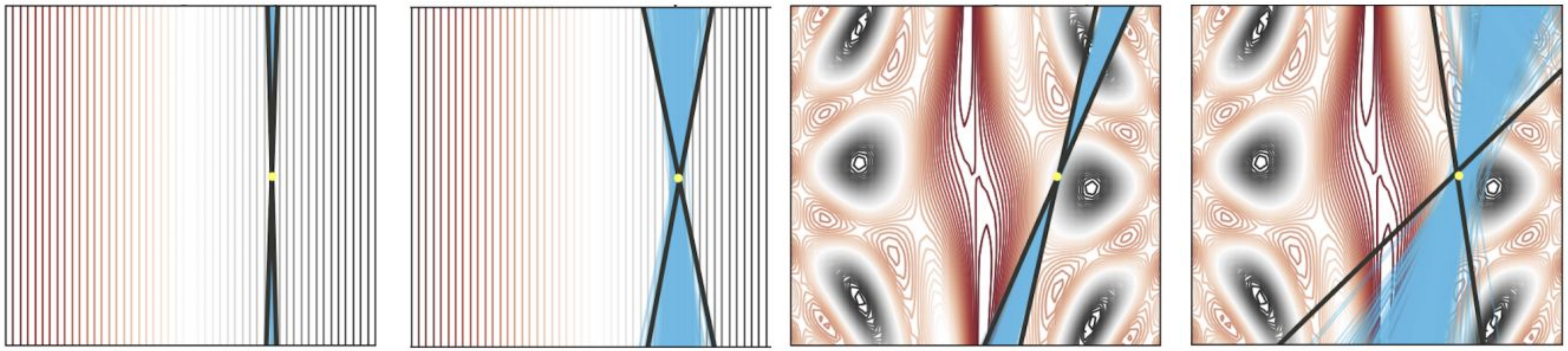
- While there are already attempts to address some of these limitations, more work is needed

Challenges with LIME: Stability

- Perturbation approaches like LIME/SHAP are unstable



Challenges with LIME: Consistency



(a) linear, many samples (b) linear, fewer samples (c) nonlinear, many samples (d) nonlinear, fewer samples

Many = 250 perturbations; Few = 25 perturbations;

When you repeatedly run LIME on the same instance,
you get different explanations (blue region)

Challenges with LIME: Consistency

Problem with having too few perturbations?

What is the optimal number of perturbations?

Can we just use a very large number of perturbations?

Challenges with LIME: Scalability

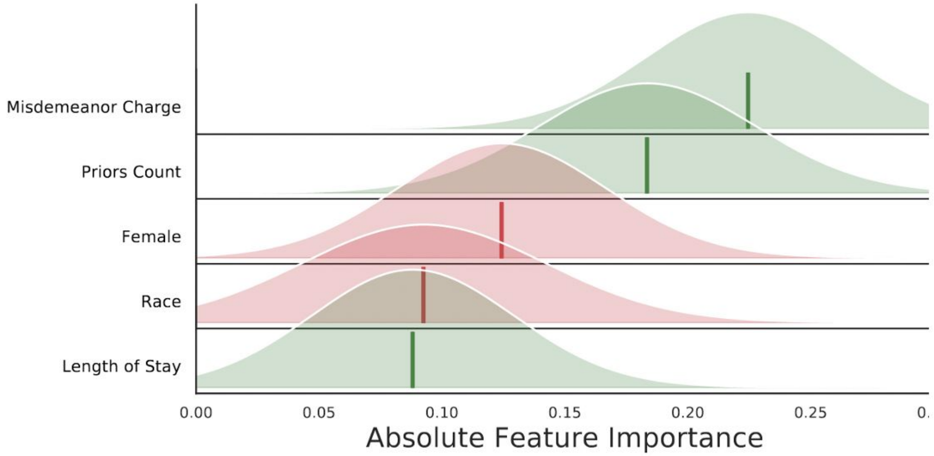
- Querying complex models (e.g., Inception Network, ResNet, AlexNet) repeatedly for labels can be computationally prohibitive

- Large number of perturbations Large number of model queries

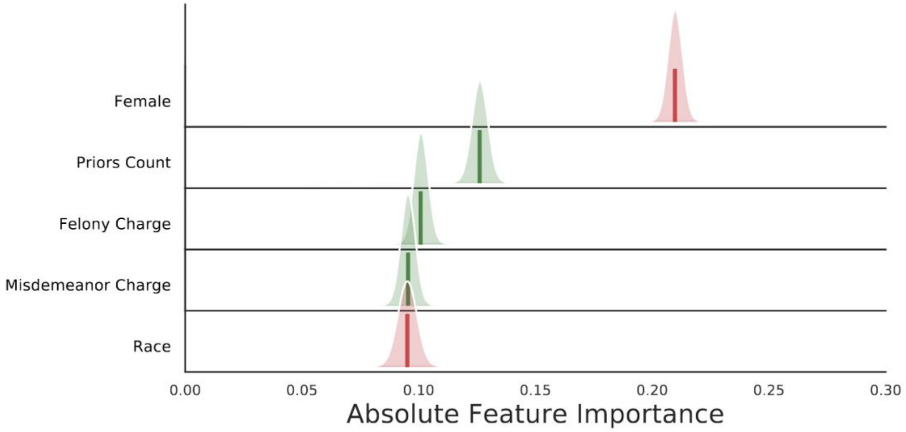
Generating reliable explanations using LIME can be computationally expensive!

Explanations with Guarantees: BayesLIME and BayesSHAP

- Intuition: Instead of point estimates of feature importances, define these as distributions



(a) Explanation computed with 100 perturbations



(b) Explanation for the same instance with 2000 perturbations

BayesLIME and BayesSHAP

- Construct a Bayesian locally weighted regression that can accommodate LIME/SHAP weighting functions

$$y|z,\phi,\epsilon \sim \phi^T z + \epsilon \qquad \epsilon \sim \mathcal{N}\left(0, \frac{\sigma^2}{\pi_x(z)}\right)$$

Black Box Predictions

Feature Importances

Perturbations

Weighting Function

$$\phi|\sigma^2 \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \qquad \sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2).$$

Priors on feature importances and feature importance uncertainty

BayesLIME and BayesSHAP: Inference

- Conjugacy results in following posteriors

$$\sigma^2 | \mathcal{Z}, Y \sim \text{Scaled-Inv-}\chi^2\left(n_0 + N, \frac{n_0\sigma_0^2 + Ns^2}{n_0 + N}\right)$$
$$\phi | \sigma^2, \mathcal{Z}, Y \sim \text{Normal}(\hat{\phi}, V_\phi \sigma^2)$$

- We can compute all parameters in closed form

These are the same equations used in LIME & SHAP!

$$\hat{\phi} = V_\phi(\mathcal{Z}^T \text{diag}(\Pi_x(\mathcal{Z}))Y)$$
$$V_\phi = \left(\mathcal{Z}^T \text{diag}(\Pi_x(\mathcal{Z}))\right) \mathcal{Z} + \mathbb{I})^{-1}$$
$$s^2 = \frac{1}{N}\left[(Y - \mathcal{Z}\hat{\phi})^T \text{diag}(\Pi_x(\mathcal{Z}))(Y - \mathcal{Z}\hat{\phi}) + \hat{\phi}^T\hat{\phi}\right]$$

Estimating the Required Number of Perturbations

I need an explanation where true feature importance lies within ±0.5 of estimated values with 95% confidence

Estimate required number of perturbations for user specified uncertainty level.

THEOREM 3.3. *Given S seed perturbations, the number of additional perturbations required (G) to achieve a credible interval width W of feature importance for a data point x at user-specified confidence level α can be computed as:*

$$G(W, \alpha, x) = \frac{4s_S^2}{\bar{\pi}_S \times \left[\frac{W}{\Phi^{-1}(\alpha)}\right]^2} - S \tag{9}$$

where $\bar{\pi}_S$ is the average proximity $\pi_x(z)$ for the S perturbations, s_S^2 is the empirical sum of squared errors (SSE) between the black box and local linear model predictions, weighted by $\pi_x(z)$, as in (7), and $\Phi^{-1}(\alpha)$ is the two-tailed inverse normal CDF at confidence level α.

Improving Efficiency: Focused Sampling

- Instead of sampling perturbations randomly and querying the black box, choose points the learning algorithm is most uncertain about and only query their labels from the black box.

This approach allows us to construct explanations with user defined levels of confidence in an efficient manner!

Other Questions

- Can we construct post hoc explanations that are provably robust to various adversarial attacks discussed earlier?

- Can we construct post hoc explanations that can guarantee faithfulness, stability, and fairness simultaneously?

Future of Model Understanding

Methods for More Reliable Post hoc Explanations

Model Understanding Beyond Classification

Theoretical Analysis of the Behavior of Interpretable Models & Explanation Methods

Intersections with Model Robustness

Empirical Evaluation of the Correctness & Utility of Model Interpretations/Explanations

Intersections with Model Fairness

Characterizing Similarities and Differences Between Various Methods

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

Theoretical Analysis of the Behavior of Explanations/Models

- We discussed some of the recent theoretical results earlier. Despite these, several important questions remain unanswered
- Can we characterize the conditions under which each post hoc explanation method (un)successfully captures the behavior of the underlying model?
- Given the properties of the underlying model, data distribution, can we theoretically determine which explanation method should be employed?
- Can we theoretically analyze the nature of the prototypes/attention weights learned by deep nets with added layers? When are these meaningful/when are they spurious?

Future of Model Understanding

- Methods for More Reliable Post hoc Explanations
- Theoretical Analysis of the Behavior of Interpretable Models & Explanation Methods
- Empirical Evaluation of the Correctness & Utility of Model Interpretations/Explanations
- Characterizing Similarities and Differences Between Various Methods

- Model Understanding Beyond Classification
- Intersections with Model Robustness
- Intersections with Model Fairness
- Intersections with Model Privacy

- New Interfaces, Tools, Benchmarks for Model Understanding

Empirical Analysis of Correctness/Utility

- While there is already a lot of work on empirical analysis of correctness/utility for post hoc explanation methods, there is still no clear characterization of which methods (if any) are correct/useful under what conditions.

- There is even less work on the empirical analysis of the correctness/utility of the interpretations generated by inherently interpretable models. For instance, are the prototypes generated by adding prototype layers correct/meaningful? Can they be leveraged in any real world applications? What about attention weights?

Future of Model Understanding

Methods for More Reliable Post hoc Explanations

Theoretical Analysis of the Behavior of Interpretable Models & Explanation Methods

Empirical Evaluation of the Correctness & Utility of Model Interpretations/Explanations

Characterizing Similarities and Differences Between Various Methods

Model Understanding Beyond Classification

Intersections with Model Robustness

Intersections with Model Fairness

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

Characterizing Similarities and Differences

- Several post hoc explanation methods exist which employ diverse algorithms and definitions of what constitutes an explanation, under what conditions do these methods generate similar outputs (e.g., top K features) ?

- Multiple interpretable models which output natural/synthetic prototypes (e.g., Li et. al, Chen et. al. etc.). When do they generate similar answers and why?

Future of Model Understanding

Methods for More Reliable Post hoc Explanations

Model Understanding Beyond Classification

Theoretical Analysis of the Behavior of Interpretable Models & Explanation Methods

Intersections with Model Robustness

Empirical Evaluation of the Correctness & Utility of Model Interpretations/Explanations

Intersections with Model Fairness

Characterizing Similarities and Differences Between Various Methods

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

[Coppens et. al., 2019, Amir et. al. 2018]

[Ying et. al., 2019]

Model Understanding Beyond Classification

- How to think about interpretability in the context of large language models and foundation models? What is even feasible here?

- Already active work on interpretability in RL and GNNs. However, very little research on analyzing the correctness/utility of these explanations.

- Given that primitive interpretable models/post hoc explanations suffer from so many limitations, how to ensure explanations for more complex models are reliable?

Future of Model Understanding

Methods for More Reliable
Post hoc Explanations

Theoretical Analysis of the Behavior of
Interpretable Models & Explanation Methods

Empirical Evaluation of the Correctness &
Utility of Model Interpretations/Explanations

Characterizing Similarities and Differences
Between Various Methods

Model Understanding
Beyond Classification

Intersections with Model Robustness

Intersections with Model Fairness

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

Intersections with Model Robustness

- Are inherently interpretable models with prototype/attention layers more robust than those without these layers? If so, why?
- Are there any inherent trade-offs between (certain kinds of) model interpretability and model robustness? Or do these aspects reinforce each other?
- Prior works show that counterfactual explanation generation algorithms output adversarial examples. What is the impact of adversarially robust models on these explanations? [Pawelczyk et. al., 2022]

Future of Model Understanding

Methods for More Reliable Post hoc Explanations

Model Understanding Beyond Classification

Theoretical Analysis of the Behavior of Interpretable Models & Explanation Methods

Intersections with Model Robustness

Empirical Evaluation of the Correctness & Utility of Model Interpretations/Explanations

Intersections with Model Fairness

Characterizing Similarities and Differences Between Various Methods

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

Intersections with Model Fairness

- It is often hypothesized that model interpretations and explanations help unearth unfairness biases of underlying models. However, there is little to no empirical research demonstrating this.

- Conducting more empirical evaluations and user studies to determine how interpetations and explanations can complement statistical notions of fairness in identifying racial/gender biases

- How does the fairness (statistical) of inherently interpretable models compare with that of vanilla models? Are there any inherent trade-offs between (certain kinds of) model interpretability and model fairness? Or do these aspects reinforce each other?

Future of Model Understanding

Methods for More Reliable
Post hoc Explanations

Model Understanding
Beyond Classification

Theoretical Analysis of the Behavior of
Interpretable Models & Explanation Methods

Intersections with Model Robustness

Empirical Evaluation of the Correctness &
Utility of Model Interpretations/Explanations

Intersections with Model Fairness

Characterizing Similarities and Differences
Between Various Methods

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

Intersections with Differential Privacy

- Model interpretations and explanations could potentially expose sensitive information from the datasets.

- Little to no research on the privacy implications of interpretable models and/or explanations. What kinds of privacy attacks (e.g., membership inference, model inversion etc.) are enabled?

- Do differentially private models help thwart these attacks?If so, under what conditions? Should we construct differentially private explanations?

Future of Model Understanding

Methods for More Reliable
Post hoc Explanations

Theoretical Analysis of the Behavior of
Interpretable Models & Explanation Methods

Empirical Evaluation of the Correctness &
Utility of Model Interpretations/Explanations

Characterizing Similarities and Differences
Between Various Methods

Model Understanding
Beyond Classification

Intersections with Model Robustness

Intersections with Model Fairness

Intersections with Model Privacy

New Interfaces, Tools, Benchmarks for Model Understanding

New Interfaces, Tools, Benchmarks for Model Understanding

- Can we construct more interactive interfaces for end users to engage with models? What would be the nature of such interactions? [demo]

- As model interpretations and explanations are employed in different settings, we need to develop new benchmarks and tools for enabling comparison of faithfulness, stability, fairness, utility of various methods. How to enable that?

Some Parting Thoughts..

- There has been renewed interest in model understanding over the past half decade, thanks to ML models being deployed in healthcare and other high-stakes settings
- As ML models continue to get increasingly complex and they continue to find more applications, the need for model understanding is only going to raise
- Lots of interesting and open problems waiting to be solved
- You can approach the field of XAI from diverse perspectives: theory, algorithms, HCI, or interdisciplinary research – there is room for everyone! ☺️

Thank You!

- **Acknowledgements**: Special thanks to Julius Adebayo, Chirag Agarwal, Shalmali Joshi, and Sameer Singh for co-developing and co-presenting sub-parts of this tutorial at NeurIPS, AAAI, and FAccT conferences.

- Email: hlakkaraju@hbs.edu; hlakkaraju@seas.harvard.edu;

- Course on interpretability and explainability: https://interpretable-ml-class.github.io/

- More tutorials on interpretability and explainability: https://explainml-tutorial.github.io/

- Trustworthy ML Initiative: https://www.trustworthyml.org/
    - Lots of resources and seminar series on topics related to explainability, fairness, adversarial robustness, differential privacy, causality etc.


TalkToModel
Hello Hima, I'm a machine learning model trained to predict whether someone has diabetes.
Let's get started. Ask me something!
Enter your command! Use the ↑ arrow and ↓ arrow to cycle previous commands.
Send
👇Help me generate a question about... 👇